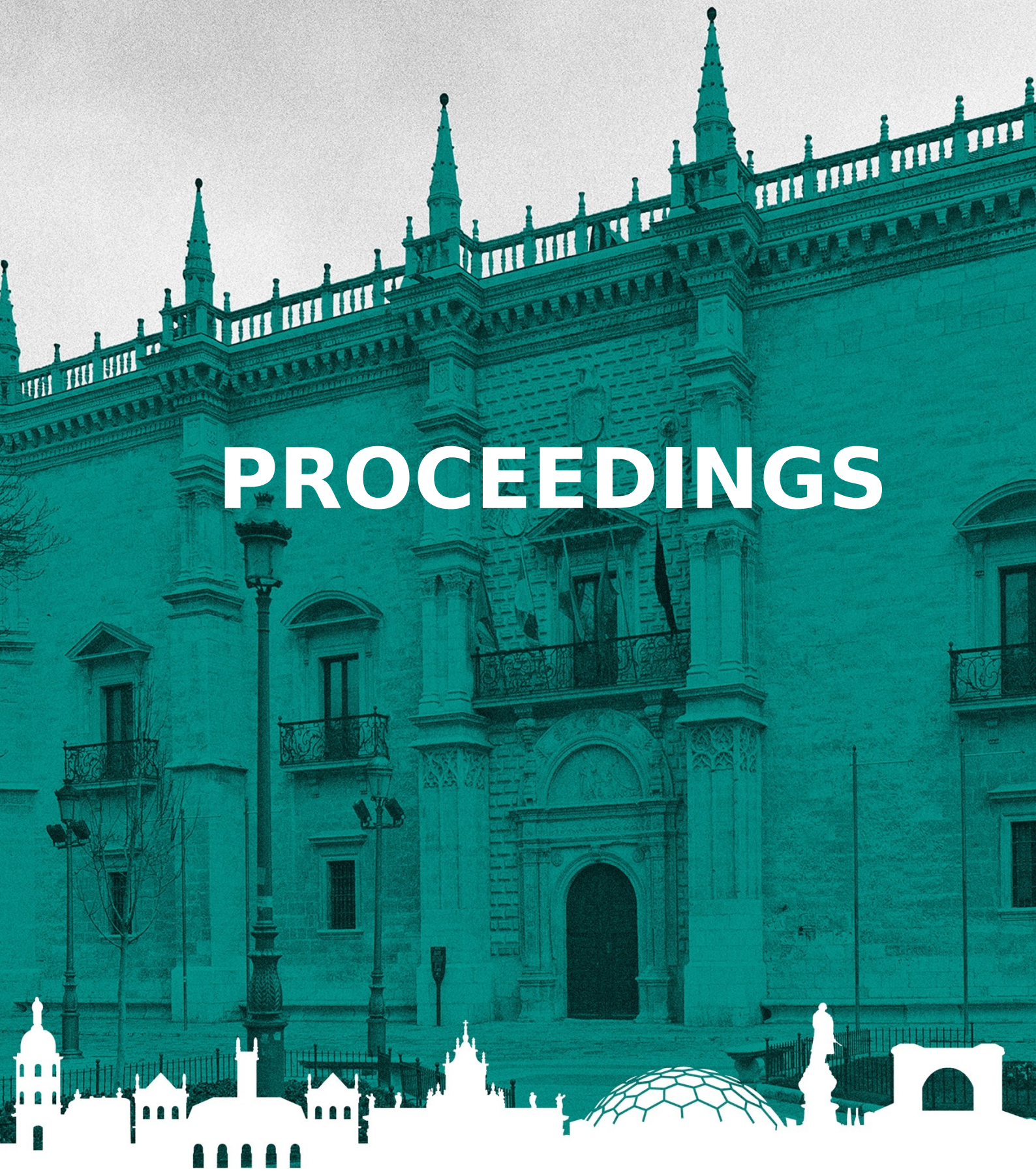
IBERSPEECH 2020

VALLADOLID | MARCH 2021



PROCEEDINGS

IBERSPEECH 2020

VALLADOLID | MARCH 2021

DOI: 10.21437/IberSPEECH.2021

Edited by Valentín Cardeñoso Payo, David Escudero Mancebo and César González Ferreras

The proceedings can be downloaded from:

IberSPEECH2020 website: https://iberspeech2020.eca-simm.uva.es
ISCA Archive: https://www.isca-speech.org/archive/IberSPEECH_2021/

Valladolid, Spain, March 2021

Welcome Message

Welcome to IberSPEECH2020 in Valladolid, March 24-25th 2021, organized by the ECA-SIMM Research group of the Department of Computer Science of the Universidad de Valladolid. The current edition has received inestimable valued support by the Spanish Thematic Network on Speech Technology (RTTH), the Cátedra RTVE and the Voice Input Voice Output Laboratory (Vivolab) at Universidad de Zaragoza, and the ISCA Special Interest Group on Iberian Languages (SIG-IL). In addition, and for the second time, IberSPEECH becomes an official ISCA Supported Event.

The IberSPEECH2020 event - the fifth of its kind using this name - brings together the 11th Jornadas en Tecnologías del Habla and the 7th Iberian SLTech Workshop events, aiming to promote interaction and discussion among junior and senior researchers in the field of speech and language processing for Iberian languages.

Valladolid is a model capital of about 300 thousand people. The historic centre of Valladolid, the city on the Pisuerga River, is home to an interesting collection of Renaissance architecture comprising houses, palaces and emblematic buildings such as the Cathedral, the College of San Gregorio (today the site of the National Sculpture Museum) and the church of San Pablo. The city has an intense cultural schedule thanks to its status as a university town, and hosts events such as the Seminci, the International Film Festival and one of the highlights of the Spanish film calendar, and the International Street Theatre and Arts Festival. Valladolid offers a unique combination of landscapes and weather, coupled with exquisite gastronomical experiences that are the result of a blend of heritage, produce, terroir, tradition, creativity, and innovation.

This year Iberspeech is organized online thanks to the support of the Palace of Conferences Conde Ansúrez. The Palace is an example of the spirit of the University of Valladolid, a building of the XVII century that has evolved until today, adapted for the different uses it has had until it is a complete infrastructure with technical media required for performing all kinds of activities, forum, workshops, conferences etc... The most recent advance in the support to online conferences as IberSPEECH 2020.

IberSPEECH2020 is a two-day event, bringing together the best researchers and practitioners in speech and language technologies in Iberian languages to promote interaction and discussion. The organizing committee has planned a wide variety of scientific and social activities, including technical paper presentations, keynote lectures, presentation of projects, laboratories activities, recent PhD thesis, discussion panels, a round table, and awards to the best thesis and papers. The program of IberSPEECH2020 includes a total of 32 contributions that will be presented distributed among 5 oral sessions, a PhD session, and a projects session. To ensure the quality of all the contributions, each submitted paper was reviewed by three members of the scientific review committee. All the papers in the conference will be accessible through the International Speech Communication Association (ISCA) Online Archive. Paper selection was based on the scores and comments provided by the scientific review committee, which includes 73 researchers from different institutions (mainly from Spain and Portugal, but also from France, Germany, Brazil, Iran, Greece, Hungary, Czech Republic, Ucrania, Slovenia). Furthermore, it is confirmed to publish an extension of selected papers as a special issue of the Journal of Applied Sciences, "IberSPEECH 2020: Speech and Language Technologies for Iberian Languages", published by MDPI with fully open access. In addition to regular paper sessions, the IberSPEECH2020 scientific program features the following activities: the ALBAYZIN evaluation challenge session.

Following the success of previous ALBAYZIN technology evaluations since 2006, this year ALBAYZIN evaluations have focused around multimedia analysis of TV broadcast content. Under the framework of the Cátedra RTVE at Universidad de Zaragoza, we introduce and report on the results of the IberSPEECH-RTVE 2020 Challenge. The Corporación de Radiotelevisión Española (RTVE) has provided participants with an annotated TV broadcast database and the necessary tools for the evaluations, promoting the fair and transparent comparison of technology in different fields related to speech and language technology. It comprises four different challenge evaluations: Speech to Text Challenge (S2TC), Speaker Diarization Challenge and Identity Assignment (SDIAC) and Multimodal Diarization and Scene Description Challenge (MDSDC), organized by RTVE and Universidad de Zaragoza; and the Search on Speech Challenge (SoSC) jointly organized by Universidad San Pablo-CEU and AuDIaS from Universidad Autónoma de Madrid with the support of the ALBAYZIN Committee. Overall, 7 teams participated in the S2TC challenge, 3 teams in the SDIAC, 3 teams in the MDC, and 2 more teams in the SoSC challenge, which resulted in 14 system paper description contributions. These were intended to describe either progress in current or recent research and development projects, demonstration systems, or PhD Thesis extended abstracts to compete in the PhD Award.

Furthermore, IberSPEECH2020 features 2 remarkable keynote speakers: Dr. Gérard Bailly (GIPSA-Lab, France) and Dr. Antonio Bonafonte (Amazon, London, UK) to whom we would like to acknowledge for their extremely valuable participation.

Valladolid, March 2021

Valentín Cardeñoso Payo, General Chair
David Escudero Mancebo, General Chair
César González Ferreras, General Chair

Organizing Committee

General Chairs
Valentín Cardeñoso Payo, Universidad de Valladolid, Spain.
David Escudero Mancebo, Universidad de Valladolid, Spain
César González Ferreras, Universidad de Valladolid, Spain

General Co-Chairs
Francesc Alías Pujol, La Salle - Universitat Ramon LLull, Spain
António Teixeira, Universidade de Aveiro, Portugal

Technical Program Chair
David Escudero Mancebo, Universidad de Valladolid, Spain

Technical Program Co-Chairs
Carlos David Martínez Hinarejos, Universitat Politècnica de València, Spain
Alberto Abad, INESC ID Lisboa / IST Lisboa, Portugal
Xavier Anguera, ELSA Corp., United States
Eva Navas, University of the Basque Country, UPV- EHU, Spain

Publication Chair
Francesc Alías Pujol, La Salle - Universitat Ramon LLull, Spain
César González Ferreras, Universidad de Valladolid, Spain

Evaluations Chair
Alfonso Ortega, Universidad de Zaragoza, Spain
Eduardo Lleida, Universidad de Zaragoza, Spain
Luis Javier Rodríguez Fuentes, Universidad del País Vasco, Spain

Local Committee
Valentín Cardeñoso Payo, Universidad de Valladolid, Spain.
David Escudero Mancebo, Universidad de Valladolid, Spain
César González Ferreras, Universidad de Valladolid, Spain
Carlos Vivaracho Pascual, Universidad de Valladolid, Spain
Mario Corrales Astorgano, Universidad de Valladolid, Spain
Cristian Tejedor García, Universidad de Valladolid, Spain

Scientific Review Committee

Alberto Abad, INESC-IST, Portugal
Francesc Alías, La Salle - Universitat Ramon Llull, Spain
Aitor Alvarez, Vicomtech-IK4, Spain
Xavier Anguera, Miro ELSA Corp., Portugal
Jorge Baptista, INESC-ID Lisboa, Portugal
Plinio Barbosa, University of Campinas, Brazil
Fernando Batista, INESC-ID & ISCTE-IUL, Portugal
José Miguel Benedí, Universitat Politècnica de València, Spain
Antonio Bonafonte, Amazon, UK.
Antonio Peinado, Universidad de Granada, Spain
José Luis Pérez Córdoba, Universidad de Granada, Spain
Ángel Gómez, Universidad de Granada, Spain
José Andrés González López, Universidad de Granada, Spain
Marcos Calvo, Google.
Joao Cabral, Trinity College Dublin, Ireland
María José Castro-Bleda, Universitat Politècnica de València, Spain
Ricardo Córdoba, Universidad Politécnica de Madrid, Spain
Conceicao Cunha, IPS Munich, Germany
Carme de-la-Mota, Universitat Autònoma de Barcelona, Spain
Arantza del Pozo, Vicomtech, Spain
Laura Docío-Fernández, University of Vigo, Spain
Daniel Erro, Cirrus Logic, Madrid, Spain
David Escudero, University of Valladolid, Spain
César González-Ferreras, University of Valladolid, Spain
Valentín Cardeñoso-Payo, University of Valladolid, Spain
Nicholas Evans, EURECOM, France
Mireia Farrús, Universitat Pompeu Fabra, Spain
Rubén Fernández, Universidad Politécnica de Madrid, Spain
Javier Ferreiros, Universidad Politécnica de Madrid, Spain
Ascensión Gallardo, Universidad Carlos III de Madrid, Spain
Fernando García Granada, Universitat Politècnica de València, Spain
Carmen García Mateo, University of Vigo, Spain
Kafentzis George, University of Crete, Greece
Omid Ghahabi, EML European Media Laboratory GmbH, Germany
Juna Ignacio Godino, Llorente Universidad Politécnica de Madrid, Spain
Jon Ander Gómez, Universitat Politècnica de València, Spain
Emilio Granell, Universitat Politècnica de València, Spain
Inma Hernaez, University of the Basque Country (UPV/EHU), Spain
Javier Hernando, Universitat Politècnica de Catalunya, Spain
Lluís-F. Hurtado, Universitat Politècnica de València, Spain
Oscar Koller, Microsoft Germany GmbH, Germany
Eduardo Lleida, University of Zaragoza, Spain
José David Lopes, Heriot Watt University UK
Paula López Otero, Universidade da Coruña, Spain
Jordi Luque, Telefónica Research, Spain
Carlos David Martínez Hinarejos, Universitat Politècnica de València, Spain
Helena Moniz, INESC/FLUL, Portugal
Juan Montero, Universidad Politécnica de Madrid, Spain

Climent Nadeu, Universitat Politècnica de Catalunya, Spain
Juan L. Navarro-Mesa, Universidad de Las Palmas de Gran Canaria, Spain
Eva Navas, University of the Basque Country, Spain
Géza Németh, Budapest University of Technology & Economics, Hungary
Nelson Neto, Universidade Federal do Pará, Brazil
Hermann Ney, RWTH Aachen University, Germany
Alfonso Ortega, University of Zaragoza, Spain
Yannis Pantazis, Foundations for Research and Technology - Hellas, Spain
Carmen Peláez-Moreno University Carlos III Madrid, Spain
Mikel Penagarikano, University of the Basque Country, Spain
Fernando Perdigao, Institute of Telecommunications (IT), Lisbon, Portugal
Ferran Pla, Universitat Politècnica de València, Spain
Jiri Pribil, Slovak Academy of Sciences Slovakia
Jorge Proenca, IT - Coimbra, Portugal
Michael Pucher, Acoustics Research Institute Austria
Paulo Quaresma, Universidade de Evora, Portugal
Ganna Raboshchuk, ELSA Corp., Portugal
Sam Ribeiro, The University of Edinburgh, UK
Eduardo Rodriguez Banga, University of Vigo, Spain
Marta Ruiz Costa-Jussà, Universitat Politècnica de Catalunya, Spain
Luis Javier Rodríguez-Fuentes, Univ. of the Basque Country UPV/EHU, Spain
Rubén San-Segundo, Universidad Politécnica de Madrid, Spain
Jon Sánchez, Aholab – EHU/UPV, Spain
Emilio Sanchis, Universitat Politècnica de València, Spain
Diana Santos, University of Oslo, Norway
Ibon Saratxaga, University of the Basque Country, Spain
Encarna Segarra, Universitat Politècnica de València, Spain
Carlos Segura Perales, Telefónica Research, Spain
Joan Serrà, Telefónica Research, Spain
Alberto Simões, 2Ai Lab - IPCA, Portugal
Rubén Solera-Ureña, INESC-ID Lisboa, Portugal
António Teixeira, University of Aveiro, Portugal
Javier Tejedor, Universidad CEU San Pablo, Spain
Doroteo Toledano, Universidad Autónoma de Madrid, Spain
Isabel Trancoso, INESC ID Lisboa / IST, Portugal
Cassia Valentini-Botinhao, The University of Edinburgh, UK
Amparo Varona, University of the Basque Country, Spain
Andrej Zgank, University of Maribor, Slovenia
Catalin Zorila, Toshiba Cambridge Research Laboratory, UK
Iván López-Espejo, Aalborg University, Denmark

Organizing Institutions

This conference has been organized by:

with the collaboration of:

IberSPEECH 2020 has been partially funded by the project Red Temática en Tecnologías del Habla 2017, (TEC2017-90829-REDT) funded by Ministerio de Ciencia, Innovación y Universidades.

Keynote 1
Characterizing and assessing the oral reading fluency of young readers

Gérard Bailly and Erika Godde

GIPSA-lab, CNRS, Grenoble Alps University

Abstract

According to the ministry of education, one young adult (16-25) over 10 has reading difficulties, 50% of them being illiterate. France was ranked 34/50 in the PIRLS 2016, last in Europe. Compared to 2011, degradation of the reading performance of these 4th grade children was also noticed. Mastering comprehensive reading is yet a perquisite for accessing other educational disciplines. But reading requires the maturation and coordination of a complex cognitive network, involving vision, phonological, semantic and pragmatic processing together with the sensorimotor activation of phonetic representations... and breathing! We will present the framework we developed so far in order to characterize, assess and improve the reading fluency of young readers. This work is performed in the context of the e-FRAN Fluence project, where hundreds of primary schoolers are trained via computer-assisted technologies and monitored in a longitudinal study.

Gérard Bailly is a senior CNRS Research Director at GIPSA-Lab, Grenoble-France of which he was deputy director (2007-2012) and headed the "Cognitive Robotics, Interactive Systems and Speech Processing" team (CRISSP). He has been working in the field of speech communication for 40 years. He supervised 33 PhD Thesis, authored 50 journal papers, 24 book chapters and more than 200 papers in major international conferences. He coedited "Talking Machines: Theories, Models and Designs" (Elsevier, 1992), "Improvements in Speech Synthesis" (Wiley, 2002) and "Audiovisual speech processing" (CUP, 2012). He is an elected member of the ISCA (International Speech Communication Association) board and a founder member of the ISCA SynSIG and SproSIG special-interest groups. His current interest is multimodal interaction with conversational agents – in particular humanoid robots – using speech, hand and head movements and eye gaze. He also works in the field of computer-aided training, in particular reading and cognitive therapy.

Erika Godde defended her PhD about reading prosody development in French children in November 2020. She is currently working at GIPSA-Lab, Université Grenoble Alpes, as a project engineer and she is looking for a postdoc. Her fields of interest are reading development from preschool to expert readers, especially reading prosody and comprehension. In broader spectrum and for future research, she is interested in langage develoepement, particularly in the development of the cognitive bases of reading prosody and comprehension, classroom applications and transfers from the lab to the classrooms.

Keynote 2
Diverse Conversational Spoken Language Generation

Antonio Bonafonte

Amazon Research, Spain

Abstract

In human speech, speaking style and prosody are often a matter of context, and for Alexa's interactions with customers to be as natural as possible, the same should be true for her. This talk presents recent work at Amazon to generate diverse and appropriated spoken responses. One of the models generates alternative phrasings in a context-aware way, so that Alexa does not keep asking the same question repeatedly. Then, the speech generator adapts the speaking style depending on the dialogue state.

Antonio Bonafonte joined Amazon TTS Research in January 2019 as Senior Researcher. The team leads the research and innovation that defines the future of Amazon Alexa and AWS Polly voices. Previously, he worked for over 25 years as Associate Professor at Universitat Politecnica de Catalunya (UPC), Barcelona, Spain, leading the TTS Research Group. Most of his career has been focussed on speech synthesis, co-authoring around 200 technical papers, and participating or leading more than 50 national or international research projects. Some of the produced technology has been included in widely available products, as the Festival Catalan voices in debian repositories, the upcTV (Catalan Educative Program) and the pioneer TaP, distributed in the 90s by CEAPAT/IMSERSO.

CONFERENCE PROGRAM

Day 1

Oral Session 1 - Applications of Speech Technologies for Learning and Education

Oral Session 2 - Speech Processing and Acoustic Event Detection

Albayzín Evaluation Challenges

Research and Development Projects

Ph.D. Thesis

Day 2

Oral Session 3 - ASR and NLP Techniques

Oral Session 4 - Speech Synthesis and Multimodal Processing

Oral Session 5 - Speaker Characterization and Diarization

275 List of Authors

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Prosodic feature selection for automatic quality assessment of oral productions in people with Down syndrome

David Escudero-Mancebo, Mario Corrales-Astorgano,
Valentín Cardeñoso-Payo, César González-Ferreras

Department of Computer Science
University of Valladolid
descuder@infor.uva.es

Abstract

Evaluation of prosodic quality is always a challenging task due to the nature of prosody with multiple form-function valid profiles. When voice of people with Down syndrome (DS) is analyzed, diversity increases making the problem even more challenging. This work is framed in our activities for developing learning games for training oral communications of people with intellectual disabilities. In this context automatic evaluation of prosodic quality is a must for deciding whether game users should repeat activities or continue playing and to inform therapists about the particular difficulties of users. In this paper we present a procedure for the selection of informative prosodic features based on both the distance between human rated right and wrong productions and the distance with respect to productions of typical users. A main contribution with respect to previous works stems from the use of mixed models to rate the impact of the type of activity and speaker dependence when estimating the quality of the prosodic productions.

Index Terms: Down syndrome speech, Computer Assisted Pronunciation Assessment.

1. Introduction

Prosody is an important component of speech communication because it is responsible for fundamental functions such as grouping linguistic units, pausing, word accent and sentence purpose (declarative, interrogative, exclamatory or imperative) and also other higher level functions like emotions and pragmatics [1]. The low control of prosody or its inappropriate production can stigmatize speakers and limit their options to get integrated in society [2]. Such could be the case of people with intellectual disabilities in general and speakers with Down syndrome (DS) in particular, which is a population characterized by special needs on language control and prosodic production (with notable exceptions) [3, 4, 5]. As far as prosody is concerned, Kent and Vorperian [4] report disfluencies (stuttering and cluttering) and impairments in the perception, imitation and spontaneous production of prosodic features; while Heselwood et al. [6] have connected some of the speech errors with difficulties in the identification of boundaries between words and sentences. In previous work we empirically showed the clear contrast between the voice of speakers with Down syndrome and typical speakers by performing perceptual and automatic identification tests from signal [7].

This work has been partially funded by Ministerio de Economía, Industria y Competitividad and the European Regional Development Fund FEDER (project TIN2017-88858-C2-1-R) and by Junta de Castilla y León (project VA050G18).

In [8] we took a step forward to analyze the possibilities to assess DS voice oral productions quality by using a similar procedure. While the problem of DS voice identification reached more than 90% accuracy with an SVM classifier [7], the problem of quality assessment reached only about 78.5% using the same type of classifier and the same training feature set. In this paper we analyze the training data used in the mentioned previous works to understand the reasons for this classification performance differences at the time that results give cues for exploring different paradigms in future works.

Software tools and learning games have been devised for intellectual disabled people to train specific competences [9, 10, 11, 12]. There are methods that voice therapists employ with speakers with specific speech problems [13]. Some of these methods, partially, have been implemented as software tools that help therapists to work with their patients or allow patients to carry out complementary exercises in an autonomous way [14]. In [15] we presented a tool to train prosody and pragmatics for speakers with DS. A set of perceptual and production activities are interleaved in a graphic adventure video game with an adapted interface that takes into account the special characteristics of individuals with Down syndrome: poor short term memory [16], attention deficits [17], problems to integrate information and deficits of language development [18]. So far, the video game has been used successfully with real users, with the assistance of an adult (the teacher, the therapist or a relative). The use of the tool has allowed the recording of a speech corpus of people with Down syndrome. The final goal of the research presented in this paper is the analysis of the potential of these recordings to train an automatic assessment system. In the medium term, this system will be integrated into the video game to allow users to train autonomously. A complete description of oral activities can be found in [19].

There are several works on automatic assessment of speech quality in atypical voices described in the literature [20, 21, 22]. However, in computer assisted pronunciation training, not only assessment is important, but also reporting information about the reasons that led the system or the expert to judge a given utterance as correctly or incorrectly produced. In [23] authors analyze how different components of speech production impact speech intelligibility in DS. In this paper we systematically analyze the prosodic features of the utterances of the corpus in order to select the most informative features and their values for predicting oral productions quality. To perform this analysis, we triangulate information obtained from the distances between right and wrong DS productions (at the glance of human evaluators) with information obtained from distances between DS productions and productions of typical users. We then impose the requirements of separation between groups and consistency.

10.21437/IberSPEECH.2021-1

In a second step, by using logistic mixed regression models, we find that features related to temporal domain are more efficient for this task than other prosodic features related to $F0$ or energy. In addition, we provide evidences that it is important to consider not only the speaker (already shown in [8]) but also the particular training activity for improving the accuracy of the automatic assessment system.

The structure of the paper is as follows. The experimental procedure section details the corpus compiled and the manual evaluation of the utterances. The procedure for the individual analysis of the different prosodic features is also presented. The results section lists the selected prosodic features and its capabilities for modeling the quality of the utterances taking into account human scores. We end the paper with a discussion that includes limitations, future work and conclusions.

2. Experimental procedure

2.1. Corpus recording

The corpus was recorded by using a graphic adventure video game [15] that requires users to perform a set of activities related to prosodic perception and production skills in order to continue playing. All the oral productions and user interactions are recorded and classified per activity and speaker while the user is playing. The current corpus has been compiled in different sessions. It has recordings of 23 speakers with Down syndrome, 966 utterances of 40 different production activities, which is about 1 hour and 10 minutes duration. More details about gender, age and metal capabilities can be found in [8]. From this work we select those 5 speakers with more than 40 utterances each (606 utterances in total), with the aim of obtaining representative results. We call this corpus θ_{DS}.

We used a second corpus of typical speakers to analyze voice of speakers with Down syndrome when contrasted with voice of typical speakers. This corpus contains a subset of the sentences recorded by the speakers with Down syndrome and has recordings of 22 speakers, with 250 utterances in total. We call this corpus θ_{TS} (details about gender and age can be found in that was used in [7])

The recordings of both corpora were made at 44,100 Hz with a Logitech PC Headset 960 USB microphone.

2.2. Human based quality evaluation

Two complementary human based evaluations have been performed:

Real-time evaluation: The production activities are assessed in real time by a therapist, seated next to the player with a secondary keyboard. The therapist can evaluate the activity as Right (θ_R) or Poor (θ_P) to let the player continue, depending of the production quality, or as Wrong (θ_W) to ask him/her to repeat the utterance. The therapist is responsible to assess the production of the gamer and the consequence is that the player has to try again to pronounce correctly until the therapist considers that he or she can continue playing. The real time scores divide the corpus: $\theta_{DS} = \theta_R \cup \theta_W \cup \theta_P$.

Off-line evaluation: Each of the utterances was rated off-line by an expert in prosody who participated in the design of the video game. The judgments were binary, corresponding to the decision of whether the utterance should be repeated or not. She declared to use the following decision criteria, adapting them to the specific activity proposed: adjustment to the expected modality (intonation); preservation of the difference between lexical stress (stressed vs. unstressed syllables) and accent (accented vs. unaccented syllables); and adjustment to the organization in prosodic groups and distinction between function and content words (phrasing). The off-line scores divide the corpus $\theta_{DS} = \theta_{R'} \cup \theta_{W'}$; R' indicating right and W' indicating wrong productions.

Both evaluations have been compared leading to consistency rates going from 79.4% to 64.1% depending on the speaker (doing R and P assignments equivalent to R')

2.3. Processing and selection of prosodic features

The openSmile toolkit [24] was used to extract acoustic features from each recording of the corpus. The GeMAPS feature set [25] was selected due to the variety of acoustic and prosodic features contained in this set: frequency related features, energy related features and temporal features. The arithmetic mean and the coefficient of variation along the utterance were calculated on these features. Furthermore, 4 additional temporal features were added: the silence and sounding percentages, silences per second and the length mean of silences. These last 4 features were calculated using the silences and sounding intervals generated by Praat software [26], which uses an intensity threshold, a minimum silent interval duration and a minimum sounding interval duration to identify these intervals (Praat default values were used). In total, 92 features were used: 10 from frequency domain, 10 from energy domain, 11 from temporal domain and 61 from spectral domain. The complete description of these features can be found in [7].

As selection criteria we require the feature f to satisfy:

1. **Separation:** there must be statistical significant differences between the values of f in the groups θ_R and θ_W (Mann-Whitney test with p-value$<$0.01) which implies that clear differences between right and wrong utterance are observed.
2. **Consistency:** being f_T, f_R and f_W the mean value of the feature f in the groups θ_{TS} (typical speakers), θ_R and θ_W respectively, it must be satisfied that $|f_T - f_R| < |f_T - f_W|$ which implies that right utterances are closer than wrong utterances to the typical ones.

We apply this procedure with both real-time and off-line evaluations. In the case of real-time evaluation the procedure is repeated for every pair of groups. The comparison of results obtained with both evaluations permits to discuss about the possible reasons for disagreement.

2.4. Analysis of the impact of features on quality

Logistic mixed effects regression models were used to measure the impact of speaker and activity on the automatic assessment of quality. This regression model takes into account that the I observations came from A different activities and S different speakers. The full model is given by

$$\begin{aligned} y_{i,a,s} = \beta_0 + A_{0,a} + S_{0,s} + \\ (\beta_\delta + A_{\delta,a} + S_{\delta,s}) X_{i,a,s} + \epsilon_{i,a,s} \end{aligned} \quad (1)$$

where β_0 is the fixed intercept and $A_{0,a}$ and $S_{0,s}$ are the random intercepts introduced by activity and speaker respectively; $A_{\delta,a}$, and $S_{\delta,s}$ are the random slopes to be added to the fixed slope β_δ. The most informative acoustic features are the fixed effect and speaker and activity are the random effects. A binomial distribution for y was used in order to build the logistic regression models. Different configurations of the mixed model

Table 1: *List of the automatically selected frequency, energy and temporal features. All the features in columns Off-line evaluation have statistically significant differences (Mann-Whitney test with p-value<0.01) between right and wrong productions of speakers with Down syndrome. The asterisk in Real-time evaluation columns means statistically significant differences (Mann-Whitney test with p-value<0.01) between first and third column (placed in the first column), between first and second column (when placed in the second column) or between the second and third column (when placed in the third column). The meaning of the features can be seen in [7]. In cells we present 95% confidence interval of the mean value. The units are reported in [24].*

			Off-line evaluation		**Real time evaluation**		
		Typical speakers	DS Right productions	DS Wrong productions	DS Right cont. prod	DS Wrong cont. prod	DS Poor productions
F0 domain							
f1	F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2	(2.40, 2.88)	(1.91, 2.67)	(2.73, 3.66)	(1.37, 2.23)*	(2.08, 3.03)	(3.48, 4.74)*
f2	jitterLocal_sma3nz_stddevNorm	(1.11, 1.21)	(1.32, 1.43)	(1.52, 1.66)	(1.35, 1.48)	(1.39, 1.56)	(1.40, 1.55)
Energy domain							
e1	loudness_sma3_percentile20.0	(0.91, 1.01)	(0.71, 0.78)	(0.63, 0.71)	(0.65, 0.72)	(0.72, 0.84)	(0.67, 0.77)
Temporal domain							
d1	loudnessPeaksPerSec	(5.64, 5.89)	(4.10, 4.32)	(3.76, 4.05)	(4.23, 4.49)*	(3.80, 4.14)*	(3.61, 3.92)
d2	StddevVoicedSegmentLengthSec	(0.14, 0.16)	(0.18, 0.23)	(0.25, 0.33)	(0.20, 0.26)	(0.20, 0.28)	(0.19, 0.27)
d3	soundingPercentage	(0.88, 0.91)	(0.89, 0.91)	(0.73, 0.79)	(0.88, 0.91)*	(0.82, 0.88)	(0.74, 0.81)*
d4	silencesPerSecond	(0.35, 0.44)	(0.28, 0.35)	(0.52, 0.63)	(0.35, 0.44)*	(0.36, 0.50)	(0.45, 0.58)
d5	silencesMean	(0.14, 0.19)	(0.13, 0.18)	(0.31, 0.41)	(0.14, 0.19)*	(0.18, 0.27)	(0.28, 0.41)

were compared in terms of the modeling capabilities with the use of the `lme4` package [27].

A reduced number of variables is used for the algorithm to iterate with 606 points, at most 3 fixed effects and 2 random effects. This procedure permits to assess both the relative importance of the acoustic features (fixed factors) and speaker and activity (random factors) on the perceived quality. We select the most informative feature of each of the three domains as fixed factors.

3. Results

Table 1 presents the 95% confidence interval of the mean value of the selected features separated by groups. Only 8 out of the 92 analyzed input features satisfy the established criteria when off-line evaluation data are contrasted (see table 1) (27 features were selected in [7] and 21 in [8]). Only 5 of them do when real-time evaluation is contrasted (asterisks in the last three columns of table 1). In fact, in real-time evaluation, when the groups DS Right vs. DS Wrong and DS Wrong vs. DS Poor are compared, only one and two variables respectively satisfy the imposed conditions (features f1, e1 and d3). This result suggests a richer evaluation in off-line conditions with more features in all the domains. In real-time evaluation the values of the energy domain variables are inconsistent in what concerns to group separation: no significant differences between groups and the closer to the typical values does not imply the more quality. Results suggest that recordings were evaluated by Poor when abnormal values of the temporal domain or F0 domain were observed (features f1, d1, d3-d5) and that utterances were marked as Wrong when speed was low (d1 feature).

Concerning the selected features, F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2 represents the range of 20-th to 80-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz and jitterLocal_sma3nz_stddevNorm represents the coefficient of variation of the deviations in individual consecutive F0 period lengths. In energy domain, loudness_sma3_percentile20.0 means the percentile 20-th of estimate of perceived signal intensity from an auditory spectrum. Finally, related with the temporal domain, loudnessPeaksPerSec means the number of the loudness peaks per second and StddevVoicedSegmentLengthSec represents the standard deviation of continuously voiced regions. soundingPercentage represents the duration percentage of voiced regions, silencesPerSecond means the number of silences per second and silencesMean represents the length mean of unvoiced regions.

The values of the confidence intervals in table 1 show that the wider the F0 range (higher f1 feature) and the less stable f0 contour (higher f2 feature) the more abnormal the utterance is perceived; the weaker the intensity (lower e1 feature) the more penalized the utterance is; utterances belonging to the wrong and poor groups are slower (lower d1 feature), have more speed changes (higher d2 feature), with more inner pauses (lower d3 and higher d4 feature) or longer pauses (higher d5 feature).

Table 1 also permits to contrast the values of features in typical vs. DS speakers. Focusing on off-line evaluation, we observe that the gap between Typical and Right productions is relevant for features f2, e1, d1, d2, d4 (with not overlapping intervals). The distance between Typical and Right utterances is higher than the distance between Right and Wrong utterances for features f2, e1 and d1. f1 and d5 features present the most overlapped intervals between Right and Typical utterances.

Table 2 shows how accurate a set of logistic regression models represent the working data. We select a feature per domain as the features in the same domain exhibit a high correlation. The incremental inclusion of new variables in the ANOVA test permits to show that all the variables significantly contribute in the modeling. The use of the variable related to duration domain (Dur in m3 and m10 models) and the inclusion of the activity in the model (m7 model), offer the most significant modelling improvements (more AIC and deviance reduction and more Acc increase).The use of random factors is a need for improving the quality of the modelling (AIC goes from 716.19 to 612.60 and Acc from 70% to 80%). Slopes of the random factors do not contribute to improve the modeling (m11 and m12 rows).

The feature selection procedure allows to identify the 8 most informative variables to predict prosodic quality from the 92 analyzed variables with satisfactory results: an automatic classifier SVM trained with the 92 variables performs with 72.0% accuracy and with the 8 variables selected performs with 71.0% (SMO -Sequential Minimal Optimization- implementation of Weka tools using normalized poly kernel with exponent

Table 2: *Summary of the sequential ANOVA test of the mixed logistic regression models when the quality of the utterance is predicted using the binary off-line evaluation. $F0$ is the variable f1 in table 1, En is the variable e1, Dur is the variable d5, $Speaker$ ranges from 1 to 5, $Activity$ ranges from 1 to 40. I means intercept and S refers to both slope and intercept.The quality metrics are the ones reported by the command anova of the package lme4 and Acc is the accuracy of the prediction of the training samples. Sig. codes: 0 '***' 0.001 '**' 0.01 '*'. AIC is the Akaike's Information Criterion [28].*

	Fixed effect			Random factor		Quality of the model					
Model	F0	En	Dur	Speaker	Activity	AIC	deviance	Chisq	Pr(>Chisq)	sig	Acc
m1	X					775.76	771.76				65%
m2		X				774.68	770.68	1.083			66%
m3			X			723.98	719.98	50.700			68%
m4	X	X				720.82	714.82	5.159	0.02313	*	69%
m5	X	X	X			716.19	708.90	5.921	0.01496	*	70%
m6	X			I		761.36	755.36				66%
m7	X				I	684.19	678.19	77.1729	$<$ 2.2e-16	***	74%
m8	X			I	I	666.17	658.17	20.0255	7.642e-06	***	75%
m9	X	X		I	I	654.38	644.38	13.7854	0.0002049	***	77%
m10	X	X	X	I	I	612.60	600.60	43.7799	3.675e-11	***	80%
m11	X	X	X	I	S	614.31	600.31	0.2872	0.5920068		80%
m12	X	X	X	S	S	616.31	600.31	0.0000	1.0000000		80%

2 [29]). Including speaker and activity information, the SVM classifier (same configuration) predicts prosodic quality with 76.8% accuracy with 8 variables and 76.7% with the 92 acoustic features (exponent 1 poly kernel in the last case).

4. Discussion

The acoustic features belonging to the temporal domain seem to be effective for the assessment of oral turns independently of speaker and activity: 5 variables out of the 8 selected features refer to temporal features (table 1) and the models including the duration of the pauses are the best predicting the quality of the utterances (table 2). This is an attractive result because the computation of this type of features, in contrast with the ones belonging to the F0 or spectral domain, is more robust in the face of the adverse conditions that could occur with users with DS. In general, the pitch detection algorithms produce more errors in pathological voices than in typical voices [30]. Furthermore, features related to the temporal domain can be easily related to disfluent speech (stuttering or cluttering) that, although not universal, is a common problem of this population [31, 32, 33].

The mixed regression model indicates that not only considering the speaker is important (already shown in [8]) but also the particular activity performed by the speaker during the recording. The specification of the particular activity was not relevant to identify whether the oral turn corresponded to a speaker with Down syndrome or not in previous works [7]. Nevertheless, here it appears to be relevant for assessing the quality of the oral turns (significantly higher precision in models m7-m10).

The search of new features, or combinations of the ones already computed, that could be related to the activity or type of activity is proposed as future work in order to improve the results, at the time that a bigger corpus is compiled for testing more sophisticated models or alternative machine learning techniques. This is a challenging task because the video game has diverse activities for players to train different language functions like asking, expressing opinions, social interaction... prosodic functions like chunking or prominence in different production modes: reading, elicited or free speech. It is present work the use of the utterances of the corpus for compiling an unsupervised classification of the activities that takes into account the different acoustic prosodic features and human judgments of quality.

5. Conclusions

The paper has presented a feature selection procedure that profits evidences from different human based evaluations and that benefits as well of empirical observations that concern with differences between Down syndrome utterances with respect to typical speakers ones.

The procedure has shown to be efficient so that 8 selected features permit predicting prosodic quality as accurately as 93 features ensemble. It permits identifying the most discriminant features per domain: temporal, frequency and energy related domain.

It has been shown and discussed the reasons why improving the accuracy of the classifier requires the consideration of the type of activity and specific profile of the user of the training tool. This fact highlights the need to implement specialized classifiers on the different type of activities and the implementation as well of user adaption techniques in future work.

6. References

[1] P. Roach, *English phonetics and phonology fourth edition: A practical course*. Ernst Klett Sprachen, 2010.

[2] B. Wells, S. Peppé, and M. Vance, "Linguistic assessment of prosody," *Linguistics in clinical practice*, pp. 234–265, 1995.

[3] R. S. Chapman and L. Hesketh, "Language, cognition, and short-term memory in individuals with Down syndrome," *Down Syndrome Research and Practice*, vol. 7, no. 1, pp. 1–7, 2001.

[4] R. D. Kent and H. K. Vorperian, "Speech impairment in Down syndrome: A review," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 1, pp. 178–210, 2013.

[5] V. Stojanovik, "Prosodic deficits in children with Down syndrome," *Journal of Neurolinguistics*, vol. 24, no. 2, pp. 145–155, 2011.

[6] B. Heselwood, M. Bray, and I. Crookston, "Juncture, rhythm and planning in the speech of an adult with down's syndrome," *Clinical Linguistics & Phonetics*, vol. 9, no. 2, pp. 121–137, 1995.

[7] M. Corrales-Astorgano, D. Escudero-Mancebo, and C. González-Ferreras, "Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome," *Speech Communication*, vol. 99, pp. 90 – 100, 2018.

[8] M. Corrales-Astorgano, P. Martínez-Castilla, D. Escudero-Mancebo, L. Aguilar, C. González-Ferreras, and V. Cardeñoso-Payo, "Automatic assessment of prosodic quality in Down syndrome: Analysis of the impact of speaker heterogeneity," *Applied Sciences*, vol. 9, no. 7, p. 1440, 2019.

[9] A. R. Cano, Á. J. García-Tejedor, C. Alonso-Fernández, and B. Fernández-Manjón, "Game analytics evidence-based evaluation of a learning game for intellectual disabled users," *IEEE Access*, vol. 7, pp. 123 820–123 829, 2019.

[10] L. E. García, R. J. Mejía, A. Salazar, and C. E. Gómez, "Un videojuego para estimular habilidades matemáticas en personas con síndrome de Down," *Revista ESPACIOS*, vol. 40, no. 05, 2019.

[11] K. Prena and J. L. Sherry, "Parental perspectives on video game genre preferences and motivations of children with Down syndrome," *Journal of Enabling Technologies*, vol. 12, no. 1, pp. 1–9, 2018.

[12] M. S. Del Rio Guerra, J. Martin-Gutierrez, R. Acevedo, and S. Salinas, "Hand gestures in virtual and augmented 3d environments for Down syndrome users," *Applied Sciences*, vol. 9, no. 13, p. 2641, 2019.

[13] D. R. Boone, S. C. McFarlane, S. L. Von Berg, and R. I. Zraick, *The voice and voice therapy*. Pearson/Allyn & Bacon Boston, 2005.

[14] W. R. Rodríguez, O. Saz, and E. Lleida, "A prelingual tool for the education of altered voices," *Speech Communication*, vol. 54, no. 5, pp. 583–600, 2012.

[15] C. González-Ferreras, D. Escudero-Mancebo, M. Corrales-Astorgano, L. Aguilar-Cuevas, and V. Flores-Lucas, "Engaging adolescents with Down syndrome in an educational video game," *International Journal of Human–Computer Interaction*, vol. 33, no. 9, pp. 693–712, 2017.

[16] R. Chapman and L. Hesketh, "Language, cognition, and short-term memory in individuals with Down syndrome," *Down Syndrome Research and Practice*, vol. 7, no. 1, pp. 1–7, 2001.

[17] M. H. Martínez, X. P. Duran, and J. N. Navarro, "Attention deficit disorder with or without hyperactivity or impulsivity in children with Down's syndrome," *International Medical Review on Down Syndrome*, vol. 15, no. 2, pp. 18–22, 2011.

[18] R. S. Chapman, "Language development in children and adolescents with Down syndrome," *Mental Retardation and Developmental Disabilities Research Reviews*, vol. 3, no. 4, pp. 307–312, 1997.

[19] D. Escudero-Mancebo, M. Corrales-Astorgano, V. Cardeñoso-Payo, L. Aguilar, C. González-Ferreras, P. Martínez-Castilla, and V. Flores-Lucas, "Prautocal corpus: A corpus for the study of down syndrome prosodic aspects." *Languages Resources and Evaluation*, vol. Under review, 2021.

[20] D. Le and E. M. Provost, "Modeling pronunciation, rhythm, and intonation for automatic assessment of speech quality in aphasia rehabilitation," in *INTERSPEECH*, 2014.

[21] M. Tu, V. Berisha, and J. Liss, "Interpretable objective assessment of dysarthric speech based on deep neural networks." in *INTERSPEECH*, 2017, pp. 1849–1853.

[22] M. Li, D. Tang, J. Zeng, T. Zhou, H. Zhu, B. Chen, and X. Zou, "An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder," *Computer Speech & Language*, vol. 56, pp. 80–94, 2019.

[23] D. O'Leary, A. Lee, C. O'Toole, and F. Gibbon, "Perceptual and acoustic evaluation of speech production in Down syndrome: A case series," *Clinical linguistics & phonetics*, pp. 1–20, 2019.

[24] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[25] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[26] P. Boersma, "Praat: doing phonetics by computer," *http://www.praat. org/*, 2006.

[27] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[28] H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[29] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burges, and A. Smola, Eds. MIT Press, 1998.

[30] S.-J. Jang, S.-H. Choi, H.-M. Kim, H.-S. Choi, and Y.-R. Yoon, "Evaluation of performance of several established pitch detection algorithms in pathological voices," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 620–623.

[31] J. Van Borsel and A. Vandermeulen, "Cluttering in Down syndrome," *Folia Phoniatrica et Logopaedica*, vol. 60, no. 6, pp. 312–317, 2008.

[32] D. Devenny and W. Silverman, "Speech dysfluency and manual specialization in Down's syndrome," *Journal of Intellectual Disability Research*, vol. 34, no. 3, pp. 253–260, 1990.

[33] K. Eggers and S. Van Eerdenbrugh, "Speech disfluencies in children with Down Syndrome," *Journal of Communication Disorders*, 2017.

Performance Comparison of Specific and General-Purpose ASR Systems for Pronunciation Assessment of Japanese Learners of Spanish

Cristian Tejedor-García[1], *Valentín Cardeñoso-Payo*[1], *David Escudero-Mancebo*[1]

[1]ECA-SIMM Research Group, Department of Computer Science, University of Valladolid, Spain
{cristian, valen, descuder}@infor.uva.es

Abstract

General–purpose state-of-the-art automatic speech recognition (ASR) systems have notably improved their quality in the last decade opening the possibility to be used in different practical applications, such as pronunciation assessment. However, the assessment of short words as minimal pairs in segmental approaches remains an important challenge for ASR, even more for non-native speakers. In this work, we use both our own tailored specific–purpose Kaldi–based ASR system and Google ASR to assess Spanish minimal pair words produced by 33 native Japanese speakers and to discuss their performance for computer-assisted pronunciation training (CAPT). Participants were split into three groups: experimental, in-classroom, and placebo. First two groups followed a pre/post-test training protocol spanning four weeks. Both the experimental and in-classroom groups achieved statistically significant differences at the end of the experiment, assessed by both ASR systems. We also found moderate correlation values between Google and Kaldi ASR systems in the pre/post-test values, and strong correlations between the post-test scores of both ASR systems and the CAPT application scores at the end of the experiment. Tailored ASR systems can bring clear benefits for a detailed study of pronunciation errors and results showed that they can be as useful as general–purpose ASR for assessing minimal pairs in CAPT tools.

Index Terms: automatic speech recognition (ASR), automatic assessment tools, foreign language pronunciation, pronunciation training, automatic pronunciation assessment, learning environments, minimal pairs

1. Introduction

Recent advances in automatic speech recognition (ASR) have made this technology a potential solution for transcribing audio input for computer-assisted pronunciation training (CAPT) tools [1, 2]. Available ASR technology, properly adapted, might help human instructors with pronunciation assessment tasks, freeing them from hours of tedious work, allowing for the simultaneous and fast assessment of several students, and providing a form of assessment that is not affected by subjectivity, emotion, fatigue, or accidental lack of concentration [3]. Thus, ASR systems can help in the assessment and feedback on learner productions, reducing human costs [4, 5]. Although most of the scarce empirical studies which include ASR technology in CAPT tools assess sentences in large portions of either reading or spontaneous speech [6, 7], the assessment of words in isolation remains a substantial challenge [8, 9].

This study has been partially supported by the Ministerio de Economía y Empresa (MINECO) and the European Regional Development Fund FEDER (TIN2014-59852-R) and by the Consejería de Educación of Junta de Castilla y León (VA050G18) and by the University of Valladolid (Ph.D. Research Grant 2015).

General–purpose off-the-shelf ASR systems like Google ASR[1] are becoming progressively popular each day due to their easy accessibility, scalability, and most importantly, effectiveness [10, 11]. These services provide accurate speech-to-text capabilities to companies and academics who might not have the possibility of training, developing, and maintaining a specific–purpose ASR system. However, despite the advantages of these systems (e.g., they are trained on large datasets and span different domains) there is an obvious need for improving their performance when used on in-domain data a specific scenarios, such as segmental approaches in CAPT for non-native speakers. Concerning the existing ASR toolkits, Kaldi has shown its leading role in recent years with its advantages of having flexible and modern code that is easy to understand, modify, and extend [12], becoming a highly matured development tool for almost any language [13, 14].

English is the most practiced language in CAPT experiments [6] and in commercial language learning applications, such as Duolingo[2] or Babbel[3]. However, there are scarce empirical experiments in the state-of-the-art which focus on pronunciation instruction and assessment for native Japanese learners of Spanish as foreign language, and as far as we are concerned, no one includes ASR technology. For instance, 1440 utterances of Japanese learners of Spanish as a foreign language (A1-A2) were analyzed manually with Praat by phonetics experts in [15]. Students performed different perception and production tasks with an instructor, and they achieved positive significant differences (at the segmental level) between the pre-test and post-test values. A pilot study on perception of Spanish stress by Japanese learners of Spanish was reported in [16]. Native and non-native participants listened to natural speech recorded by a native Spanish speaker and were asked to mark one of three possibilities (the same word with three stress variants) of an answer sheet. Non-native speech was manually transcribed with Praat by phonetic experts in [17], in an attempt to establish rule-based strategies for labeling intermediate realizations, helping to detect both canonical and erroneous realizations in a potential error detection system. Different perception tasks were carried out in [18]. It was reported how the speakers of native language (L1) Japanese tend to perceive Spanish /y/ when it is pronounced by native speakers of Spanish; and how the L1 Spanish and L1 Japanese listeners evaluate and accept various consonants as allophones of Spanish /y/, comparing both groups.

In previous work, we presented the development and the first pilot test of a CAPT application with ASR and text-to-speech technology, Japañol, through a training protocol [19, 20]. This learning application for smart devices includes a specific exposure–perception-production cycle of training activities

[1]https://cloud.google.com/speech-to-text
[2]https://www.duolingo.com/
[3]https://www.babbel.com/

10.21437/IberSPEECH.2021-2

with minimal pairs which are presented to students in lessons of the most difficult Spanish contrasts for native Japanese speakers. We were able to empirically measure statistically significant improvement between the pre and post-test values of 8 native Japanese speakers in a single experimental group. The students' utterances were assessed by experts in phonetics and by Google ASR system, obtaining strong correlations between human and machine values. After this first pilot test, we wanted to take a step further and to find pronunciation mistakes associated with key features of proficiency level characterization of more participants (33) and different groups (3). However, assessing such a quantity of utterances by human raters derived to a problem of time and resources. Also, Google ASR pricing policy and its limited black-box functionality also motivated us to look for alternatives to assess all the utterances, developing an own ASR system with Kaldi. In this work, we analyze the audio utterances of the pre-test and post-test of 33 Japanese learners of Spanish as foreign language with two different ASR systems (Google and Kaldi) to address the question of how these general and specific–purpose ASR systems can deal with the assessment of short words in the field of CAPT.

This paper is organized as follows. The experimental procedure is described in section 2, which includes the participants and protocol definition, a brief description about the process for elaborating the Kaldi–based ASR system, and the collection of metrics and instruments for collecting the necessary data. Results section shows the word error rate (WER) values of the Kaldi–based ASR system developed, the pronunciation assessment of the participants at the beginning and at the end of the experiment, including intra and inter-group differences, and the ASR scores' correlation of both ASR systems. We end this paper with a discussion about the performance of both state-of-the-art ASR systems in CAPT supported by our results and we shed light on lines of future work.

2. Experimental Procedure

2.1. Participants

A total of 33 native Japanese speakers from 18 to 26 years old participated voluntarily for the experimental prototype. All of them declared a low level of Spanish as foreign language with no previous training in Spanish phonetics. None of them stayed in any Spanish speaking country for more than 3 months. Besides, they were requested not to do any extra work in Spanish (e.g., conversation exchanges with natives or extra phonetics research) while the experiment was still active.

Participants were randomly divided into three groups: (1) **experimental group**, 18 students (15 female, 3 male) who trained their Spanish pronunciation with Japañol, during three sessions of 60 minutes; (2) **in-classroom group**, 8 female students who attended three 60-minutes pronunciation teaching sessions within the Spanish course, with their usual instructor, making no use of any computer-assisted interactive tools; and (3) **placebo group**, 7 female students who only took the pre-test and post-test. They did not attend neither the classroom nor the laboratory for Spanish phonetics instruction.

Finally, a group of 10 native Spanish speakers from the theater company Pie Izquierdo of Valladolid (5 women and 5 men) participated in the recording of a total of 41,000 utterances (7.1 hours of speech data) for the training corpus of the Kaldi ASR system for assessing the students' utterances gathered during the experimentation.

2.2. Protocol Description

We followed a four-week protocol which included a pre-test, three training sessions, and a post-test for the non-native participants. Native speakers recorded the speech training corpus for the Kaldi–based ASR system. At the beginning, the non-native subjects took part in the pre-test session individually in a quiet testing room. The utterances were recorded with a microphone and an audio recorder (the procedure was the same for the post-test). All the students took the pre-test under the sole supervision of a member of the research team. They were asked to read aloud the 28 minimal pairs administered via a sheet of paper with no time limitation[4]. The pairs came from 7 contrasts of the most difficult to perceive and produce Spanish consonant sounds by native Japanese speakers (see more details in [19]): [θ]–[f], [θ]–[s], [fu]–[xu], [l]–[ɾ], [l]–[r], [ɾ]–[rr], and [fl]–[fr]. Students were free to repeat each contrast as many times as they want if they thought they might have mispronounced them. Each participant took an average of 83.77 seconds to complete the pre-test (63.85 seconds min. and 129 seconds max.) and an average of 94.10 seconds to complete the post-test (52.45 and 138.87 seconds min. and max.).

From the same 7 contrasts, a total of 84 minimal pairs[4] were presented to the experimental and in-classroom group participants in 7 lessons along three training sessions. The minimal pairs were carefully selected by experts taking into account the Google ASR limitations (homophones, word-frequency, very short words, and out-of-context words, in a similar process as in [8]). The lessons were included in the CAPT tool for the experimental group and during the class sessions for the in-classroom group (12 minimal pairs per lesson, 2 lessons per session, except for the last session that included 3 lessons, see more details about the training activities in [19]). The training protocol sessions were carried out during students course's classes, in which a minimal pair was practiced in each lesson (blocked practice) and most phonemes were retaken in later sessions (spaced practice). Regarding the sounds practiced in each session, in the first one, sounds [fu]–[xu] and [l]–[ɾ] were contrasted. In the second one, [l]–[r] and [ɾ]–[rr]. The last session involved the sounds [fl]–[fr], [θ]–[f], and [θ]–[s]. Finally, subjects of the placebo group did not participate in the training sessions. They were supposed to take the pre-test and post-test and obtain results without significant differences. All participants were awarded with a diploma and a reward after completing all stages of the experiment.

On the other hand, each one of the native speakers recorded individually 164 words[4] for 25 times (41,000 utterances in total) presented randomly in five-hour sessions, for elaborating the training corpus for the Kaldi–based ASR system. The average, minimum, maximum, and standard deviation of the words length were: 4.29, 2, 8, and 1.07, respectively. The phoneme frequency (%) was: [a]: 16.9, [o]: 11.3, [r]: 9.0, [e]: 7.8, [f]: 5.3, [s]: 5.0, [ɾ]: 4.8, [l]: 4.5, [t]: 3.6, [k]: 3.6, [u]: 3.2, [i]: 3.2, [θ]: 3.2, [n]: 2.8, [m]: 2.3, [ɣ]: 1.8, [j]: 1.4, [ð]: 1.5, [x]: 1.3, [b]: 1.3, [p]: 1.1, [d]: 1.1, [β]: 0.9, [w]: 0.9, [ŋ]: 0.7, [g]: 0.3, [ʤ]: 0.2, and [z]: 0.1. The recording sessions were carried out in an anechoic chamber of the University of Valladolid with the help of a member of the ECA-SIMM research group.

2.3. Elaborating an ASR System with Kaldi

We analyzed the pre/post-test utterances of the participants with Kaldi and Google ASR systems. We did not have access to

[4] https://github.com/eca-simm/minimal-pairs-japanol-eses-jpjp

enough human resources to carry out the perceptual assessment of such a quantity of audio files, and Google ASR system just offered a limited black-box functionality and specification, so that, we developed our in–house Kaldi–based ASR system. In order to do so, different phoneme-level train models were tested in the Kaldi ASR system with the audio dataset recorded with native speakers before assessing the non-native test utterances.

The ASR pipeline that we have implemented uses a standard Gaussian Mixture Model-Hidden Markov Model (GMM/HMM) architecture, adapted from existing Kaldi recipes [12, 21]. After collecting and preparing the speech data for training and testing, the first step is to extract acoustic features from the audio utterances and training monophone models. To train a model, monophone GMMs are first iteratively trained and used to generate a basic alignment. Triphone GMMs are then trained to take surrounding phonetic context into account, in addition to clustering of triphones to combat sparsity. The triphone models are used to generate alignments, which are then used for learning acoustic feature transforms on a per-speaker basis in order to make them more suited to speakers in other datasets [22]. In our case, we re-aligned and re-trained these models four times (tri4).

2.4. Instruments and Metrics

We gathered data from five different sources: (1) a registration form with student's demographic information, (2) pre-test utterances, (3) log files and (4) utterances of user's interaction with Japañol, and (5) post-test utterances. Personal information included name, age, gender, L1, academic level, and final consent to analyze all gathered data. Log files gathered all low-level interaction events with the CAPT tool and monitored all user activities with timestamps. From these files we computed a CAPT score per speaker which refers to the final performance at the end of the experiment. It includes the number of correct answers in both perception and production (in which we used Google ASR) tasks while training with Japañol [19]. Pre/post-test utterances consisted in oral productions of the minimal pairs lists provided to the students.

A set of experimental variables was computed: (1) WER values of the train/test set models for the specific–purpose Kaldi ASR system developed in a [0, 100] scale; (2) the student's pronunciation improvement at the segmental level comparing the difference of number of correct words at the beginning (pre-test) and at the end (post-test) of the experiment in a [0, 10] scale. We used this scale for helping teachers to understand the score as they use it in the course's exams. This value consists on the mean of correct productions in relation to the total of utterances. Finally, (3) the correlation values between Google and Kaldi ASR systems of the pre/post-test utterances and between the CAPT score and both ASR systems at the end of the experiment (post-test) in a [0, 1] scale.

By way of statistical metrics and indexes, Wilcoxon signed–rank tests have been used to compare the differences between the pre/post-test utterances of each group (intra-group), Mann–Whitney U tests have been used to compare the differences between the groups (inter-group), and Pearson correlations have been used to explain the statistical relationship between the values of the ASR systems and the final CAPT scores.

3. Results

Table 1 shows the models tested for native (Kaldi) and non-native (Kaldi and Google) speech data gathered, and the WER value reported by Google for natives [10]. Regarding the native models, the *All* model included 41,000 utterances of the native speakers in the train set. The *Female* model included 20,500 utterances of the 5 female native speakers in the train set. The *Male* model included 20,500 utterances of the 5 male native speakers in the train set. The *Best1*, *Best2*, and *Best3* models included 32,800 utterances (80%) of the total of native speakers (4 females and 4 males) in the train set. These last three models were obtained by comparing the WER values of all possible 80%/20% combinations (train/test sets) of the native speakers (e.g., 4 female and 4 male native speakers for training: 80%, and 1 female and 1 male for testing: 20%), and choosing the best three WER values (the lowest ones). On the other hand, the non-native test model consisted of 3,696 utterances (33 participants x 28 minimal pairs x 2 words per minimal pair x 2 tests).

Table 1: *WER values (%) of the ASR systems.*

	Train model						
		Kaldi					
	Google	*All*	*Female*	*Male*	*Best1*	*Best2*	*Best3*
Native	5.0	0.0024	3.10	1.55	0.14	0.14	0.23
Non-native	30.0	44.22	55.91	64.12	46.40	46.98	48.08

Google reported a 5.0% WER for their English ASR system for native speech [10]. Their training techniques are applied also for their ASR in other majority languages, such as Spanish. Thus, we extrapolated this WER value to our Spanish experiment. Regarding the Kaldi-based ASR system, we achieved values lower than 5.0% for native speech for the specific battery of minimal pairs introduced in Section 2 (e.g., *All* model: 0.0024%). On the other hand, we tested the non-native minimal pairs utterances with Google ASR obtaining a 30.00% (16% non-recognized words). In the case of the Kaldi–based ASR, as expected, the *All* model reported the best test results (44.22%) for the non-native speech. The *Female* train model derived into a better WER value for the non-native test model (55.91%) than the *Male* one (64.12%) since 30 out of 33 participants were female speakers.

Table 2 shows the mean scores assigned by the Google and Kaldi ASR systems to the 3,696 utterances of the pre/post-tests classified by the three groups of participants, in a [0, 10] scale. The students who trained with the tool (experimental group) achieved the best pronunciation improvement values in both Google (0.7) and Kaldi (1.1) ASR systems. However, the in-classroom group achieved better results in both tests and by both ASR systems (4.1 and 6.1 in the post-test; and 3.5 and 5.2 in the pre-test, Google and Kaldi, respectively). The placebo group achieved the worst post-test (3.2 and 3.5, Google and Kaldi, respectively) and pronunciation improvement values (0.2 and 0.4, Google and Kaldi, respectively).

A Wilcoxon signed–rank test found statistically significant intra-group differences between the pre- and post-test values of the experimental and in-classroom groups of both ASR systems. In the case of the placebo group, there were differences only in the Google ASR values (see p and Z values in Table 2). Concerning inter-group pairs comparisons, a Mann–Whitney U test found statistically significant differences between the experimental and in-classroom groups in the post-test Google ASR scores ($p < 0.001$; Z = -2.773) and Kaldi ones ($p < 0.001$; Z = -2.886). There were also differences between the experimental and placebo groups in the post-test Kaldi scores ($p < 0.001$;

Table 2: *Pre/post-test scores. $\overline{n}$, N, and Δ refer to mean score of the correct pre/post-test utterances, number of utterances, and difference between the post and pre-test mean scores, respectively.*

	Pre-test				Post-test				Δ (Post-test - Pre-test) – Wilcoxon signed–rank test					
	Google		Kaldi		Google		Kaldi		Google			Kaldi		
Group	$\overline{n}$	N	$\overline{n}$	N	$\overline{n}$	N	$\overline{n}$	N	Δ	p-value	Z	Δ	p-value	Z
Experimental	3.0	560	4.1	560	3.7	560	5.2	560	0.7	< 0.001	-13.784	1.1	< 0.001	-5.448
In-classroom	3.5	448	5.2	448	4.1	448	6.1	448	0.6	< 0.001	-2.888	0.9	< 0.001	-3.992
Placebo	3.0	392	3.1	392	3.2	392	3.5	392	0.2	0.002	-3.154	0.4	0.059	-1.891

Z = -5.324). Post-test differences between the in-classroom and placebo groups were only found in the Kaldi scores ($p < 0.001$; Z = -7.651). Finally, although there were significant differences between the pre-test scores of the in-classroom group and the experimental group (Google: $p < 0.001$; Z = -8.892; Kaldi: $p < 0.001$; Z = -3.645), and the placebo group (Google: $p < 0.001$; Z = -8.050; Kaldi: $p = 0.001$; Z = -3.431), such differences were minimal since the effect size values were small (r = 0.10 and r = 0.20, respectively).

Table 3: *Regression coefficients of the ASR and CAPT systems. x, y, a, b, $S.E.$, and r refer to dependent variable, independent variable, slope of the line, intercept of the line, standard error, and Pearson coefficient, respectively.*

x	y	a	b	$S.E.$	r	p-value
pre-Kaldi	pre-Google	0.927	1.919	0.333	0.51	0.005
post-Kaldi	post-Google	0.934	1.897	0.283	0.57	0.002
post-Google	CAPT	0.575	-0.553	0.148	0.81	0.002
post-Kaldi	CAPT	0.982	-1.713	0.314	0.74	0.007

Finally, we analyzed several correlations between (1) the pre/post-test scores of both ASR systems (three groups) and (2) the CAPT scores with the experimental group post-test scores of both ASR systems (only group with a CAPT score) in order to compare the three sources of objective scoring (Table 3). The first and second rows of Table 3 represent the moderate positive Pearson correlations found between the Google and Kaldi pre-test (r = 0.51, p = 0.005) and post-test (r = 0.57, p = 0.002) scores. Finally, the third and fourth rows of Table 3 represent the fairly strong positive Pearson correlations found between the CAPT scores and the post-test scores of Google (r = 0.81, p = 0.002) and Kaldi (r = 0.74, p = 0.007) ASR systems.

4. Discussion and Conclusions

We have reported on empirical evidences about significant pronunciation improvement at the segmental level of native Japanese beginner–level speakers of Spanish by using state-of-the-art ASR systems (Table 2). In particular, the experimental and in-classroom group speakers improved 0.7|1.1 and 0.6|0.9 points out of 10, assessed by Google|Kaldi ASR systems, respectively, after just three one-hour training sessions. These results agreed with those reported in [8, 23]. Thus, the training protocol and the technology included, such as the CAPT tool and the ASR systems provided a very useful and didactic instrument that can be used complementary with other forms of second language acquisition in larger and more ambitious language learning projects.

Our specific–purpose Kaldi ASR system allowed us to reliably measure the pronunciation quality of the substantial quantity of utterances recorded. In particular, this ASR system proved to be useful for working at the segmental (phone) level for non-native speakers. Developing an in–house ASR system allowed us not only to customize the post-analysis of the speech without the black-box and pricing limitations of the general–purpose Google ASR system, but also neither pre-discard specific words (e.g., infrequent, out-of-context, and very short words) nor worry about the data privacy. Despite the positive results reported about the Kaldi ASR, the training corpus was limited in both quantity and variety of words and the experiment was carried out under a controlled environment. Noise-reduction, data augmentation, and a systematic study of the non-native speech data gathered to find pronunciation mistakes associated with key features of proficiency level characterization with the help of experts for its automatic characterization [4, 17] must be considered in the future to expand the project.

We have also compared our Kaldi ASR results with Google ASR ones, obtaining moderate correlations between them (Table 3). Although our specific–purpose ASR system is neither as accurate nor ambitious as Google ASR, it seems to be promising and robust enough for a concrete battery of minimal pairs. Hence, both state-of-the-art ASR systems proved to be valid for our pronunciation assessment task of minimal pair words. Future work will consist on a fine–tuning of our Kaldi–based ASR system with more utterances and re-training techniques, such as deep or recurrent neural networks, combining both native and non-native speech in order to improve current results and to obtain a better customization of the ASR system to the specific phone–level tasks. Thus, researchers, scholars, and developers can decide which one to integrate in their CAPT tools depending on the tasks and resources available.

Finally, the post-test values of both Google and Kaldi ASR systems strongly correlated with the final scores provided by the CAPT tool of the experimental group speakers (Table 3). That is, although the training words in Japañol were not the same as the pre/post-test ones, the phonemes trained were actually the same and the speakers were able to assimilate the lessons learned from the training sessions to the final post-test. Therefore, we were able to ensure that both scoring alternatives are valid and can be used for assessing Spanish minimal pairs for certain phonemes and contexts (e.g., resources availability, learning, place, data privacy, or costs).

5. Acknowledgements

The authors would like to thank Mr. Takuya Kimura for his support, the participants of the University of Seisen (Japan) and Language Learning Center of University of Valladolid, and the theater company Pie Izquierdo of Valladolid.

6. References

[1] E. Martín-Monje, I. Elorza, and B. G. Riaza, *Technology-Enhanced Language Learning for Specialized Domains: Practical Applications and Mobility*. Oxon, UK: Routledge, 2016. [Online]. Available: https://doi.org/10.4324/9781315651729

[2] O'Brien *et al.*, "Directions for the future of technology in pronunciation research and teaching," *J. Second Lang. Pronunciation*, vol. 4, no. 2, pp. 182–207, Feb. 2018. [Online]. Available: https://doi.org/10.1075/jslp.17001.obr

[3] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy-technology interface in computer-assisted pronunciation training," *Comput. Assisted Lang. Learn.*, vol. 15, no. 5, pp. 441–467, Aug. 2010. [Online]. Available: https://doi.org/10.1076/call.15.5.441.13473

[4] J. v. Doremalen, C. Cucchiarini, and H. Strik, "Automatic pronunciation error detection in non-native speech: The case of vowel errors in Dutch," *J. Acoustical Soc. America*, vol. 134, no. 2, pp. 1336–1347, 2013. [Online]. Available: https://doi.org/10.1121/1.4813304

[5] T. Lee *et al.*, "Automatic speech recognition for acoustical analysis and assessment of Cantonese pathological voice and speech," in *Proc. ICASSP*, Shanghai, China, Mar. 20-25, 2016, pp. 6475–6479. [Online]. Available: https://doi.org/10.1109/ICASSP.2016.7472924

[6] R. I. Thomson and T. M. Derwing, "The effectiveness of L2 pronunciation instruction: A narrative review," *Appl. Linguistics*, vol. 36, no. 3, pp. 326–344, Jul. 2015. [Online]. Available: https://doi.org/10.1093/applin/amu076

[7] G. Seed and J. Xu, "Integrating technology with language assessment: Automated speaking assessment," in *Proc. ALTE*, Bologna, Italy, May 3–5, 2017, pp. 286–291.

[8] C. Tejedor-García, D. Escudero-Mancebo, E. Cámara-Arenas, C. González-Ferreras, and V. Cardeñoso-Payo, "Assessing pronunciation improvement in students of English using a controlled computer-assisted pronunciation tool," *IEEE Trans. Learn. Technol.*, vol. 13, no. 2, pp. 269–282, Mar. 2020. [Online]. Available: https://doi.org/10.1109/TLT.2020.2980261

[9] J. Cheng, "Real-time scoring of an oral reading assessment on mobile devices," in *Proc. Interspeech*, Hyderabad, India, Sep. 2–6, 2018, pp. 1621–1625. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-34

[10] M. Meeker, "Internet trends 2017," may 2017, Kleiner Perkins, Los Angeles, CA, USA, Rep. [Online]. Available: https://www.bondcap.com/report/it17.

[11] V. Këpuska and G. Bohouta, "Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx)," *Int. J. Eng. Res. Appl*, vol. 7, no. 03, pp. 20–24, 2017. [Online]. Available: https://doi.org/10.9790/9622-0703022024

[12] D. Povey *et. al*, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, Waikoloa, Hawaii, HI, USA, Dec. 11–15, 2011, pp. 1–4.

[13] P. Upadhyaya, S. K. Mittal, O. Farooq, Y. V. Varshney, and M. R. Abidi, "Continuous Hindi Speech Recognition Using Kaldi ASR Based on Deep Neural Network," in *Mach. Intell. Signal Anal.*, M. Tanveer and R. B. Pachori, Eds. Singapore: Springer Singapore, 2019, pp. 303–311. [Online]. Available: https://doi.org/10.1007/978-981-13-0923-6_26

[14] I. Kipyatkova and A. Karpov, "DNN-Based Acoustic Modeling for Russian Speech Recognition Using Kaldi," in *Speech Comput.*, A. Ronzhin, R. Potapova, and G. Németh, Eds. Cham: Springer International Publishing, 2016, pp. 246–253. [Online]. Available: https://doi.org/10.1007/978-3-319-43958-7_29

[15] G. F. Lázaro, M. F. Alonso, and K. Takuya, "Corrección de errores de pronunciación para estudiantes japoneses de español como lengua extranjera," *Cuadernos CANELA*, vol. 27, pp. 65–86, Jan. 2016.

[16] T. Kimura, H. Sensui, M. Takasawa, A. Toyomaru, and J. J. Atria, "A Pilot Study on Perception of Spanish Stress by Japanese Learners of Spanish," in *Proc. SLATE*, Tokyo, Japan, Sep. 22–24, 2010.

[17] M. Carranza, "Intermediate phonetic realizations in a Japanese accented L2 Spanish corpus," in *Proc. SLaTE*, Grenoble, France, Aug./Sep. 30–1, 2013, pp. 168–171.

[18] T. Kimura and T. Arai, "Categorical Perception of Spanish /y/ by Native Speakers of Japanese and Subjective Evaluation of Various Realizations of /y/ by Native Speakers of Spanish," *Speech Res.*, vol. 23, pp. 119–129, 2019.

[19] C. Tejedor-García, V. Cardeñoso-Payo, M. J. Machuca, D. Escudero-Mancebo, A. Ríos, and T. Kimura, "Improving Pronunciation of Spanish as a Foreign Language for L1 Japanese Speakers with Japañol CAPT Tool," in *Proc. IberSPEECH*, Barcelona, Spain, Nov. 21–23, 2018, pp. 97–101. [Online]. Available: https://doi.org/10.21437/IberSPEECH.2018-21

[20] C. Tejedor-García, V. Cardeñoso-Payo, and D. Escudero-Mancebo, "Japañol: a mobile application to help improving Spanish pronunciation by Japanese native speakers," in *Proc. IberSPEECH*, Barcelona, Spain, Nov. 2018, pp. 157–158.

[21] E. Chodroff, "Corpus Phonetics Tutorial," 2018. [Online]. Available: https://arxiv.org/abs/1811.05553

[22] D. Povey and G. Saon, "Feature and model space speaker adaptation with full covariance Gaussians," in *Proc. Interspeech*, Pittsburgh, PA, USA, Sep. 17–21, 2006, pp. 1145–1148.

[23] C. Tejedor-García, D. Escudero-Mancebo, V. Cardeñoso-Payo, and C. González-Ferreras, "Using challenges to enhance a learning game for pronunciation training of English as a second language," *IEEE Access*, vol. 8, no. 1, pp. 74 250–74 266, Apr. 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.2988406

An ASR-based Reading Tutor for Practicing Reading Skills in the First Grade: Improving Performance through Threshold Adjustment

Yu Bai[1], *Ferdy Hubers*[1,2], *Catia Cucchiarini*[1], *Helmer Strik*[1,2,3,4]

[1] Centre for Language and Speech Technology (CLST), Radboud University Nijmegen
[2] Centre for Language Studies (CLS), Radboud University Nijmegen
[3] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen
[4] NovoLearning B.V., Nijmegen

y.bai@let.ru.nl, f.hubers@let.ru.nl, c.cucchiarini@let.ru.nl, w.strik@let.ru.nl

Abstract

Automatic Speech Recognition (ASR) technology can potentially be employed to provide intensive practice and feedback to young children learning to read. So far there has been limited research on the use of ASR in the early stages of learning to read when children are still developing decoding skills. For this purpose, we developed an ASR-based system equipped with logging capabilities that can evaluate decoding skills in Dutch first graders reading aloud and provide them with instantaneous feedback. In a previous study we found that ASR-based feedback led to improved reading accuracy and speed, and that useful information could be obtained from the log-files. For the present paper we conducted thorough analyses of the data obtained with this ASR-based system by comparing it to human annotations of the same read aloud 11849 words from 38 pupils. We present the results of our analyses, and discuss how they can contribute to better and more personalized ASR-based reading instruction.

Index Terms: Automatic Speech Recognition, reading tutor, child speech

1. Introduction

Automatic Speech Recognition (ASR) technology has previously been incorporated in education software for learning to read, because of its potential to provide intensive practice in reading aloud and, possibly, to detect reading problems. As a consequence, ASR technology has been mostly used to follow children while they read aloud a text and to identify upcoming reading difficulties, for instance because children hesitate to read specific words. In turn, support could be provided to teach pupils the correct form by resorting to text-to-speech technology. In our own research we decided to investigate the usability of ASR at earlier stages of learning to read, when children are still acquiring decoding skills. Within the framework of the DART project (https://dart.ruhosting.nl/), we developed an ASR-based system equipped with logging capabilities that can evaluate whether Dutch first graders reading aloud read words and sentences correctly. In turn the system provides feedback and opportunities for rehearsal so that pupils can practice and improve both reading accuracy and speed. In a previous study [1] we found that the feedback provided by our system managed to improve both reading accuracy and speed. In addition, the log-files provided useful insights that could be employed to improve both practice and feedback. One aspect that was not investigated in that study was the performance of the ASR-based system in establishing whether words were read correctly or incorrectly. In the present paper we address this issue by comparing the performance of our system to annotations of the same read aloud speech obtained from 38 pupils. The aim is to gain insights into the performance of our system with a view to improving it, for instance by applying different thresholds at which a word is judged to be incorrect. In addition, these more detailed analyses can provide more specific information on which letters and sounds are more problematic for children to read and for the ASR to recognize, respectively. Both types of insights can increase our understanding of the nature of reading errors and ASR errors and can thus contribute to developing better and more personalized ASR-based reading instruction.

2. Research Background

Several studies have investigated the contribution of ASR technology in supporting children learning to read. Most of the research addressed English reading skills [2]–[6], while studies on other languages and Dutch in particular have been limited [7]–[11]. Most of the commercial applications that have been developed so far aim at monitoring children while they read aloud and at providing support when they hesitate, for example because they do not know how a word should be read aloud. For English, commercial products exist that employ ASR in this way like the Reading Assistant (http://www.readingassistant.com/), IBM Reading Companion (https://www.ibm.com/ibm/responsibility/downloads/initiatives/ReadingCompanion.pdf), and the ReadingBuddy (http://readingbuddysoftware.com/#).

However, we think that ASR could provide a valuable contribution in earlier stages of learning to read, when children are developing decoding skills [12], that is when they learn how to convert letters to speech sounds, and need to focus on both accuracy and fluency [13]. In these stages ASR can be employed to determine whether children manage to read words and sentences correctly [2]–[4], but this requires dedicated algorithms that can identify reading errors at more detailed levels. In addition, if feedback has to be provided instantaneously, then the decision whether the word was read correctly or not also has to be instantaneous and it has to be based on single observations, which, of course, is quite challenging. To develop this kind of system it is extremely important to gain insight into ASR performance to determine to what extent these decisions are made appropriately by the system. To get an idea of the challenging nature of making such decisions, it is also important to note that even human raters often disagree on what should be considered a reading error

10.21437/IberSPEECH.2021-3

[14]. So far there has been limited research that has compared ASR performance to human performance at this level of detail in an online system, which is understandable in view of the time-consuming nature of the annotations required. The methodology applied in the current study, which is partly similar to the one described in [1], is presented below.

3. Method

3.1. Reading tutor

The reading tutor was developed to follow as much as possible the reading method for first graders developed by Zwijsen Publishers, 'Veilig Leren Lezen', which is the most widely used reading method in Dutch primary schools [15]. Two types of exercises were incorporated which address different reading skills: accuracy exercises and fluency exercises. The accuracy exercises focus on the pupils' reading accuracy of individual words and sentences. The pupil clicks on the recording button and reads one word or sentence. With the ASR backend giving scores on each word, the reading tutor gives feedback on whether the target word or sentence is correct. If the target word or sentence is read incorrectly, the pupil has to try again (up to a maximum of three attempts).

The fluency exercises aim to improve the pupils' reading fluency while keeping track of accuracy at the same time. In the fluency exercises, pupils practice reading word lists and stories. They are instructed to read a word list or a story in one go (the first try). Subsequently, they receive feedback and are prompted to try the incorrect words or sentences again. Next, pupils have to read the same word list or story again (the second try) and are encouraged to read faster with the same accuracy.

See [1] for a detailed description of the system and the reading materials used in the system.

3.2. ASR Technology

The reading tutor uses the NovoLearning [https://www.novo-learning.com/] ASR engine, which analyses the spoken attempts by the children by calculating scores (probabilities) at the phone and word level. These scores are expressed in numbers ranging from 0 to 100. The score at the word level is the minimum score of all the phones the word consists of, and is used to provide feedback. If this score is lower than the threshold, the child gets feedback that the word was read incorrectly. For the current data collection, the threshold was set to 50. All scores are stored in log-files, together with other relevant information such as the onset and offset of the speech, and the number of attempts. The audio and log-files are stored.

3.3. Data Collection

38 Dutch pupils from Grade 1 in six primary schools together read 28543 words. The pupils were between 6 and 7 years old and were in the early stages of learning to read. The experiment was conducted during the COVID-19 pandemic, so most pupils used the reading tutor at home, in an uncontrolled context.

3.4. Manual Transcription and DP alignment

Out of the 28543 words, 12185 words (42.69%) were orthographically transcribed by a human annotator who was familiar with the procedure adopted in Dutch primary schools to score tests of reading accuracy and fluency. From these words, 4095 words were taken from the accuracy exercises and 8090 words from the fluency exercises. From the accuracy exercises, we selected words with a third attempt (meaning that the first two attempts by the same pupil were judged to be incorrect, as explained in Section 3.1). The first and second attempt of that specific word were also included. To balance the number of correct and incorrect words judged by the software, the same words that were judged as correct at the first and second attempts were also selected for transcription. From the fluency exercises, word lists and stories at the first and second try were selected. To balance the number of correct and incorrect words, the same words that were judged as correct in the fluency exercises were selected as well. The transcriber made orthographic transcriptions of words coming from all four types of exercises. The procedure was as follows: First, Praat [16] textgrids containing the prompt were automatically generated using a Python script. Next, the transcriber used Praat to verify whether the prompt had actually been read. If this was the case, she did not have to change anything and could move on to the next textgrid. If the prompt was not a correct orthographic transcription of what had been read, the transcriber was instructed to change the orthographic transcription in line with what she had heard.

Figure 1: *The procedure of DP alignment.*

After the manual transcription had been obtained, we used the dynamic programming algorithm ADAPT [17] to align the manual transcription with the prompt. Figure 1 presents an example of how the prompt and manual transcription were aligned and how the words were judged to be correct based on the manual transcription. It is important to note that this process may not be straightforward in certain cases. For instance, a child may read the words a couple of times and sometimes the correct word is produced only at the end of the utterance (see Figure 1). In such cases it is important that the ASR detects the correct word in the whole utterance, because this is the way teachers normally judge such reading attempts. In establishing the degree of correspondence between the manual transcription and the prompt this should be also taken into account. For example, if the reader tried to read a word multiple times, the last attempt should be aligned to the word in the prompt. To achieve this, we used the ADAPT algorithm in inverse direction and aligned the prompt and the manual transcription from right to left instead of from left to right (STEP 1). In this way, the words in the prompt were aligned with the last occurrence of the word in the manual transcription, which means that the last trial of a word in the prompt was recognized. Words in green in Figure 1 are pronounced correctly in the last trial. For the words that

were not aligned (words in red in Figure 1), we used the normal ADAPT without reversion so that more words were aligned (STEP 2). The 'FINAL RESULT' is a binary score for each word: 0 – incorrect, 1 – correct. [Bo is a boy's name.]

3.5. Data Analysis

To gain insight in the performance of our system, we compared the system-based judgements with the human-based judgements of word reading by applying different thresholds. To this end, we calculated different measures for thresholds varying from 0 to 100. Cohen's kappa was calculated as a measure of agreement between the system-based and the human-based judgements [18], and we also calculated measures such as the percentages of correct acceptance (CA), correct rejection (CR), false acceptance (FA) and false rejection (FR), precision (P), recall (R), and the F-measure as in [19] to get insight into the quality of the feedback. This terminology and related measures were preferred because they are more transparent than those indicating false positives, negatives, etc., which can vary depending on the perspective from which the binary classification is analyzed [20].

4. Results

In this section, we present a general overview of the results concerning the ASR scoring (Section 4.1) and then go on to investigate how optimal thresholds can be established to improve the performance of the ASR system (Section 4.2).

4.1. General results of ASR scoring

12185 words read by 38 pupils were transcribed. For 2.8% of these words, no transcription could be provided, because the corresponding audio recordings were empty. We excluded these words so that in total 11849 words remained for the analysis. Figure 2 shows the frequency count of the word probability scores given by the ASR backend in the system. The majority of words received a high word probability, meaning that it is highly likely that these words were read correctly.

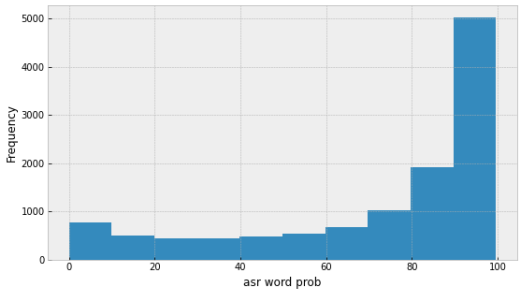


Figure 2: *Distribution of the word probabilities.*

4.2. Optimal ASR thresholds

Figure 3 shows Cohen's kappa as a function of the threshold ranging from 0 to 100. The optimal threshold is 48. Using this threshold gives the highest agreement between the system-based judgements and the human-based judgements. Cohen's kappa at a threshold of 48 is .41, which indicates moderate agreement [21].

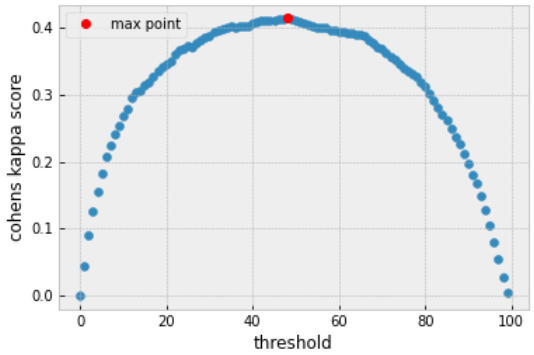


Figure 3: *Cohen's kappa for different thresholds.*

Based on the word probability scores, the reading tutor classifies words into correct and incorrect with a threshold that is manually set from 0 to 100. This classification results in four outcomes: correct acceptance (CA), correct rejection (CR), false acceptance (FA) and false rejection (FR). Figure 4 shows the percentages of CA, CR, FA and FR for thresholds from 0 to 100. At a threshold of 36, the percentages of false rejects, correct rejects and false accepts are about equal. The percentage of correct accepts is very high in general, but seems to drop rapidly after a threshold of about 70.

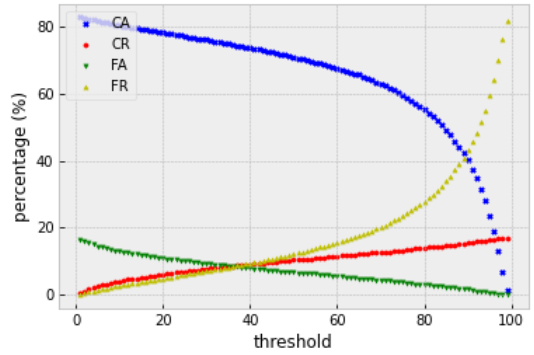


Figure 4: *CA, CR, FA and FR for different thresholds.*

Figure 5 shows F-measures of CA and CR for thresholds from 0 to 100. The threshold with the highest F-measure of CR (52.51%) is 48, which is close to the optimal threshold based on Cohen's kappa and which leads to an F-measure of CA that is still above 80%.

Table 1a shows the percentages of CA, CR, FA and FR for different thresholds from low to high. We included two extreme thresholds (20 and 60), a threshold of 36 at which the error rate (FA and FR) is equal, the optimal threshold based on Cohen's kappa and the F-measure for CA (48), and the threshold that we used in our data collection (50).

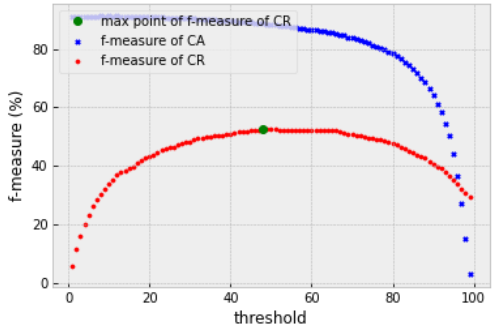


Figure 5: *F-measures of CA and CR as a function of different thresholds.*

As can be expected, we see in this table that when the applied threshold is stricter the percentage of CA decreases, while the percentage of CR increases. In addition, the percentages of FA and FR are generally low. Even with a threshold of 60, the percentage of false rejects is far below 20%.

We calculated the mean probability scores for each pupil who read over 300 words. An example of a poor reader with a mean probability of 54.796% and an example of a good reader with a mean probability of 84.257% were selected. The maximum Cohen's kappa (0.349) for the poor reader is at a threshold of 35, while the maximum Cohen's kappa (0.371) for the good reader is at a threshold of 56.

Table 1: *CA, CR, FA and FR at different thresholds, (a) average for all readers, (b) poor, (c) good reader*

Readers	**Thresholds**	**CA (%)**	**CR (%)**	**FA (%)**	**FR (%)**
		(a)			
All readers	20	78.4	6.0	10.9	4.8
	36	74,6	8.5	8.4	8.4
	48	71.5	10.1	6.8	11.6
	50	70.9	10.3	6.6	12.2
	60	67.4	11.4	5.5	15.6
		(b)			
Poor reader	20	62.7	11.5	15.3	10.5
	36	57.0	15.7	11.1	16.2
	48	51.1	17.7	9.1	22.1
	50	49.4	17.9	8.9	23.8
	60	40.6	19.1	7.8	32.5
		(c)			
Good reader	20	88.5	1.6	8.0	2.0
	36	87.4	2.7	6.9	3.0
	48	85.7	3.6	6.0	4.7
	50	85.6	3.7	5.9	4.8
	60	83.9	4.4	5.2	6.6

Table 1b and Table 1c show the percentages CA, CR, FA and FR of the poor reader and the good reader at different thresholds. When comparing these values, it can be seen that changing the threshold especially leads to substantial changes in all measures for a poor reader. For example, raising the threshold from 20 to 60, increases the percentage of FR by 22%. For a good reader, however, the measures do not change much as the threshold increases. Here changing the threshold from 20 to 60 only results in a 4.6% increase of FR.

5. Discussion and Conclusions

Within the DART project, we developed an ASR-based reading tutor that provides instantaneous feedback on the correctness of words and sentences read aloud by first graders, to be used in carefully controlled experiments in schools. Because of COVID 19 these experiments could not be conducted and children were allowed to use the system at home, in an uncontrolled context. An important requirement of the reading tutor was that it should ignore initial, often incomplete, attempts at reading the words, and should evaluate only the last attempt. The results show that the reading tutor was capable of doing this. However, there is room for improvement, for instance, by improving the acoustic models and language model of the ASR, and by studying how to obtain an optimal score for the whole word (now we use the minimum probability of all the phones composing the word).

In the current paper we investigated the effect of applying different thresholds on the performance of this online reading tutor. The results show that for all our data together, the maximum value of Cohen's Kappa and F(CR) are at a threshold of 48, which is very close to the threshold of 50 which was determined in the pilot phase and eventually adopted in the experiment proper. Increasing the threshold makes the system stricter and leads to higher values of CR and FR and lower CA and FA. Between 40 and 60, the changes are small and above 70 large changes are observed. In terms of feedback errors, the balance between FA and FR is important, as there is a trade-off between these two types of errors. Missing errors (FA) leads to reduced corrective feedback and possibly to less effective feedback. Flagging correctly read words as being incorrect (FR) can cause frustration and demotivation. In our informal observations during the pilot experiments, we noticed that the latter was indeed the case. So a good strategy in choosing the threshold probably is to prioritize reducing FR. The performance obtained in our experiment is comparable to that achieved in previous research on Dutch ASR-based reading assessment [11], with the important difference that that was a corpus-based study, while the present research refers to an online system used in an uncontrolled environment. Deciding what the optimal threshold should be in practice, will depend also on many other factors, such as the proficiency level of the child (see Table 1), the difficulty level of the content, the goals of the system, and preferences of the teachers and the pupils. An important final observation is that although the performance of the current system can still be enhanced, this was in any case sufficient to bring about improvement, as we saw in [1], and children were already positive about the current system.

6. Acknowledgements

We are grateful to Anna Krispin for transcribing the data and to Wieke Harmsen for helping us with the ADAPT alignment and the automated scoring. In addition, we would like the thank our colleagues Marjoke Bakker and Erik van Schooten as well as our partners in the DART project [http://dart.ruhosting.nl/]: NovoLearning [https:// www.novo-learning.com/], esp. Joost van Doremalen and David van Leeuwen; and Zwijsen publishers [https://www.zwijsen.nl/], esp. Rosemarie Irausquin and Martin de Jong. Special thanks go to all the children who participated, their parents and their teachers. This work (project number 40.5.18540.121) is funded by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), the Dutch Organization for Scientific Research.

7. References

[1] Y. Bai, F. Hubers, C. Cucchiarini, and H. Strik, "ASR-Based Evaluation and Feedback for Individualized Reading Practice," in *Interspeech 2020*, 2020, pp. 3870–3874.

[2] J. Mostow, J. Nelson-Taylor, and J. E. Beck, "Computer-Guided Oral Reading versus Independent Practice: Comparison of Sustained Silent Reading to an Automated Reading Tutor That Listens," *J. Educ. Comput. Res.*, vol. 49, no. 2, pp. 249–276, Sep. 2013.

[3] K. Reeder, J. Shapiro, J. Wakefield, and R. D'Silva, "Speech Recognition Software Contributes to Reading Development for Young Learners of English," *Int. J. Comput. Lang. Learn. Teach.*, vol. 5, no. 3, pp. 60–74, Aug. 2015.

[4] B. Wise, R. Cole, S. Van Vuuren, S. Schwartz, L. Snyder, N. Ngampatipatpong, and J. Tuantranont, "Learning to Read with a Virtual Tutor: Foundations to Literacy," in *Interactive Literacy Education: Facilitating literacy learning environments through technology*, C. Kinzer and L. Verhoeven, Eds. Mahwah, NJ: Lawrence Erlbaum, 2005.

[5] A. Alwan, Y. Bai, M. Black, L. Casey, M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman, and S. Wang, "A system for technology based assessment of language and literacy in young children: The role of multiple information sources," in *2007 IEEE 9Th International Workshop on Multimedia Signal Processing, MMSP 2007 - Proceedings*, 2007, pp. 26–30.

[6] M. P. Black, J. Tepperman, and S. S. Narayanan, "Automatic prediction of children's reading ability for high-level literacy assessment," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 4, pp. 1015–1028, 2011.

[7] L. Cleuren, "Elements of Speech Technology Based Reading Assessment and Intervention," Ph.D dissertation, KU Leuven, Belgium, 2009.

[8] J. Duchateau, L. Cleuren, H. Van Hamme, and P. Ghesquière, "Automatic assessment of children's reading level," in *Interspeech 2007*, 2007, pp. 1210–1213.

[9] J. Duchateau, Y. O. Kong, L. Cleuren, L. Latacz, J. Roelens, A. Samir, K. Demuynck, P. Ghesquière, W. Verhelst, and H. Van Hamme, "Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules," *Speech Commun.*, vol. 51, no. 10, pp. 985–994, Oct. 2009.

[10] M. Nicolao, M. Sanders, and T. Hain, "Improved Acoustic Modelling For Automatic Literacy Assessment Of Children," in *Interspeech 2018*, 2018.

[11] E. Yilmaz, J. Pelemans, and H. Van Hamme, "Automatic assessment of children's reading with the FLaVoR decoding using a phone confusion model," in *Interspeech 2014*, 2014, pp. 969–972.

[12] A. Castles, K. Rastle, and K. Nation, "Ending the Reading Wars: Reading Acquisition From Novice to Expert," *Psychol. Sci. Public Interes.*, vol. 19, no. 1, pp. 5–51, 2018.

[13] J. J. Pikulski and D. J. Chard, "Fluency: Bridge Between Decoding and Reading Comprehension," *Read. Teach.*, vol. 58, no. 6, pp. 510–519, 2005.

[14] P. Price, J. Tepperman, M. Iseli, T. Duong, M. Black, S. Wang, C. K. Boscardin, M. Heritage, P. David Pearson, S. Narayanan, and A. Alwan, "Assessment of emerging reading skills in young native speakers and language learners," *Speech Commun.*, vol. 51, no. 10, pp. 968–984, Oct. 2009.

[15] M. J. C. Mommers, L. Verhoeven, and S. Van der Linden, *Veilig Leren Lezen*. Tilburg: Zwijsen, 1990.

[16] P. Boersma and D. Weenink, Praat: doing phonetics by computer [Computer program]. Version 6.1.38, retrieved 30 December 2020 from http://www.praat.org/.

[17] B. Elffers, C. van Bael, and H. Strik, "ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions," Internal report, University of Nijmegen, 2005.

[18] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[19] S. Kanters, C. Cucchiarini, and H. Strik, "The goodness of pronunciation algorithm: a detailed performance study.," in *Speech & Language Technology in Education -SLaTE*, 2009, no. 2, pp. 2–5.

[20] D. M. W. Powers, "Evaluation : From Precision , Recall and F-Factor to ROC , Informedness , Markedness & Correlation," Adelaide, Australia, 2007.

[21] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, p. 159, 1977.

Impact of Vowel Reduction in L2 chinese learners of Portuguese

Catarina Realinho[1], *Rita Gonçalves*[1], *Helena Moniz*[1,2], *Isabel Trancoso*[2,3]

[1]FLUL/CLUL, Universidade de Lisboa, Lisboa, Portugal
[2]Human Language Technologies, INESC-ID Lisboa, Lisboa, Portugal
[3]Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal

cmrrafael@campus.ul.pt, ritamgg@gmail.com, helenam,imt@hlt.inesc-id.pt

Abstract

Connected speech processes have a great impact on word recognition. This is particularly impactful for L2 learners. Vowel reduction is a very productive process in European Portuguese and plays an important role within and across word boundaries. Our goal is to understand the influence of these phonetic-phonological phenomena in L2 Chinese learners of European Portuguese in word identification tasks, in continuous speech, in a classroom environment. We designed a perception experiment involving these phenomena in increasing degrees of difficulty: isolated word identification without (i) and with vowel reduction (ii); word identification with simple (iii) and complex connected speech processes (iv). This study took place in a classroom setting and it was applied to a group of B1 students of European Portuguese and a group of native speakers (control group). Data from L2 oral productions were also collected to compare it with perception data. The rate of correct answers for each task matched our expectations: (i) 94%; (ii) 65%; (iii) 31%; (iv) 16%. The results reveal that word recognition is challenging for L2 learners due to the connected speech processes. Vowel reduction brings difficulties for learners, even when the word is isolated. Production analysis reveals that learners do not produce vowel reduction, nor any other phonetic-phonological phenomena across word boundaries.

Index Terms: Second Language; Phonology; Perception; Production; Vowel Reduction; Connected Speech Processes

1. Introduction

Vowel reduction within and across word boundaries is usually described as a very frequent process in European Portuguese (EP). In continuous speech in informal contexts, it can be so extreme that may even be perceived by a non-native European Portuguese speaker as a Slavic language, due to consonant clusters. L2 perception studies usually use lab-based input. However, in (semi)spontaneous speech, sounds are altered, deleted, contracted, and this may affect L2 speech perception. The interplay between vowel reduction and connected speech processes in both perception and production tasks is not studied in a context of L2 learners of EP. To this end, in this paper we describe the impact of these phonetic-phonological phenomena in L2 Chinese learners of EP. The experiments were conducted in an ecological setting of an intensive Portuguese course, intermediate level (B1), at the University of Lisbon. A control group of native speakers also performed the experiments. The main goals of this study are threefold: (i) to describe the impact of vowel reduction and continuous speech processes in perception tasks; (ii) to analyze the presence/absence of vowel reduction phenomena in production tasks; (iii) to understand the interplay between the complexity of the linguistic structure and vowel reduction patterns, at the intermediate level (B1). We started by designing a series of perception experiments with increasing degrees of difficulty: from isolated word identification, without and with vowel reduction, to word identification in continuous speech with simple and complex contexts across word boundaries. The production data from students was also analyzed, to map the presence/absence of vowel reduction or/and other phonetic- phonological phenomena with the oral competences pertaining to that intermediate level. This paper is structured as follows: Section 2 summarizes the related work; Section 3 presents the experimental design of our perception and production tasks; Section 4 analyzes the perception results, whereas Section 5 deals with the production data. Finally, Section 6 presents our conclusions and future work.

2. Related work

Listening in second language is a complex process, as extensive literature has been pointing out. Efficient L2 listening is described with similar phases to native listening: segmenting continuous speech, distinguishing phoneme contrasts, activating words from vocabulary and dealing with other phonetic-phonological phenomena to achieve a successful listening[1]. An L2 learner is biased by his/her L1, which acts as filter conditioning both perception and production. Speakers use language-specific strategies in speech segmentation and word recognition [2], [3]. Thus, problems arise when both L1 and L2 are phonologically distant [4], [5], [6]. The vocabulary of an L2 learner is much smaller than his/her L1 vocabulary. Thus, it may trigger issues in L2 vocabulary activation and may even cause a competition between words and "non-words" activation [7]. Furthermore, the form of a spoken word may be altered due to connected speech processes[8], [9]. As a consequence, learners may recognize spoken isolated words but may not recognize that same words in continuous speech [2]. Thus, segmenting continuous speech requires knowledge of word-boundary patterns and distinct phonological inventories may trigger a wrong segmentation [10], [11], [12]. Moreover, meaning is a strong cue in word recognition. Research on this topic also revealed prosody plays an important role in detecting L2 syntactic boundaries [13].

In our research, the two languages are very distinct in terms of phonological rules, [14], [15], [16]. In European Portuguese, stress position may vary within the word [17] triggering lexical contrasts [18]. In addition, vowel reduction within and across word boundaries [17] and connected speech processes [19] are very productive in EP. The interplay of these phonological-phonetic phenomena results in the production of multiple consonant clusters [16] which has a great impact in the EP rhythmic patterns. In contrast, Mandarin Chinese (MC) has lexical tones [20] and a simple syllabic structure [14], [15] allowing the occurrence of only one segment at the onset position.The resulting

10.21437/IberSPEECH.2021-4

absence of consonant clusters contributes to a syllable-timed rhythmic structure [21]. Thus, Mandarin speakers may apply these patterns when listening to European Portuguese utterances [1] or when producing Portuguese utterances [13], which may be perceived and assessed as a non-fluent strategy.

Portuguese research on this topic [22] is related to spoken interlanguage rhythm of L2 Chinese learners of EP regarding an intermediate (B1) and an advanced level (C2). In the reading tasks, B1 students produced more vowel segments, contrasting with more advanced speakers (C2) who lowered the vowel production, resembling a fluent native Portuguese speaker, but still not in the same proportion of a native [16], [17]. This proves that at intermediate or advanced levels, speakers still transfer properties from their L1 to the L2 [13], [1].

A final note on speech data and its ecological usages. Second language teaching should provide sufficient training in order to improve students' perception, production, and comprehension skills in the target language [13], [23], [24]. L2 teaching studies usually focus on learners' errors, describing what kind of errors they produce [13]. If teachers have a clear understanding of the common difficulties that students experience [3], it will contribute to develop or adapt teaching strategies and didactic materials [2]. For instance, teaching students of how some words and sounds change in spontaneous speech may encourage their ability to segment continuous speech [8]. Thus, second language teaching must assure students are exposed to authentic speech materials and real communicative contexts [9], [23], [25], [1]. It is, therefore, important to include listening exercises with real interviews, for instance, and perform classroom activities or games to encourage oral interactions between students and teacher [24]. This was the bases o four work and the main distinction between what was scarcely been produced for L2 EP learning. We tackle read and spontaneous speech in an ecological classroom environment with exercises mimicking part of what the teacher would conduct with the class.

3. Experimental scenarios

3.1. Participants

To understand the acquisition trajectory of the phonetic processes, we selected a group of students at the intermediate level(B1) of EP (according to the Common European Framework of Reference for Languages (CEFR)). B1 students can understand the general message and the main points of clearly articulated standard speech, as in TV programmes, discussions, lectures, etc., provided the topics are familiar to them.

The class consists of Mandarin Chinese native speakers (N=12), aged between 20 and 25. Only a student also spoke Cantonese. Previously to the course, the students learned Portuguese in China and then took an intensive Portuguese course at the University of Lisbon. They were in Portugal approximately since the beginning of the course. In addition, a group of EP native speakers (N=12), aged between 23 and 27, also performed the experiments in order to set a baseline. They were either undergraduate or postgraduate students at the University of Lisbon.

3.2. Perceptual experiments

The time frame and the order of the experiments were planned so as to study the impact of vowel reduction in perception tasks, and simultaneously reinforce the didactic outcomes of the results. The experiments were integrated into the topics taught in the course, so that the experiments could be seen as a natural extension of the classroom dynamics.

Table 1: *Examples of words in Task 1.*

Word	**Translation**	**Transcription**
Identificação	Identification	[idẽtifikɐ'sɐ̃w̃]
Promoção	Promotion	[pɾumu'sɐ̃w̃]
Telefone	Telephone	[tɨle'fɔnɨ]

Table 2: *Examples of words in Task 2.*

Word	**Translation**	**Transcription**
At**i**vidade	Activity	[ɐtvi'dad]
Prod**u**tores	Producers	[pɾud'toɾʃ]
Entr**e**vista	Interview	[ẽtr'viʃtɐ]

3.2.1. Single word identification tasks

The first experiment consisted of two single words identification tasks. The first task encompassed 10 prosodic (stressed) words, all nouns, with an extension of 4 syllables on average (see Table 1, for examples). To ensure students knew the words, they were all selected from exercises performed within the same teaching unit, in their classes. The stimuli were previously recorded by a female European Portuguese native speaker, in a silent room, using the recording system from a Samsung Galaxy Note II smartphone. Recordings were converted to WAV format and edited in Praat software[26]. All words were produced without using vowel reduction. Students heard the stimuli of each isolated word, and wrote it down as soon as they identified it. Each word was presented twice. The task took around 10 minutes to be completed. The experiment took place in approximately a month after the beginning of the course.

The second task, similarly to the previous one, was also single word identification, in which students were exposed to a sequence of 10 isolated words, but this time pronounced with vowel reduction. This task was performed in the same day, and took 10 minutes as well. The vowel reduction patterns closely followed the frequent ones described for EP, as established in [17], such as the deletion of [i], [u] and [ɨ] in unstressed positions. Again, all words were selected from previous exercises (see Table 2 for examples). Recording and edition processes were replicated.

3.2.2. Cross word identification tasks with vowel reduction and other phonetic-phonological phenomena

The second experiment was more challenging than the first one, by means of combining vowel reduction with other very productive continuous speech processes in EP. For this reason, it was essential to use authentic speech material presenting real communicative contexts. This experiment (8 minutes) replicates common exercises students performed in the classroom. For the first task, we selected a YouTube news video, 01:29 minutes long, related to the topics the students were learning. Only the audio of the video was presented (twice). The students were given a form with an incomplete orthographic transcription, from which 9 word pairs were missing (totaling 18 different words). The selected missing words that the students had to fill in were prosodic words from varied lexical categories, with 3 syllables on average.The words were familiar to the students,

Table 3: *Examples of word sequences in Task 3.*

Word	Translation	Transcription
nove anos	nine years	['nɔv'ɐnuʃ]
já estava	was already	['ʒaʃ'tavɐ]
jovens saem	youngster leave	['ʒovéjʃ'saɹɐ̃j̃]
países preferidos	preferred countries	[pɐ'izʃpɾf'ɾidʃ]
dois mil	two thousand	['doɹʒ'miɻ]
depois vêm	then come	[d'poɹʒ'vɐ̃j̃ɐ̃j̃]
melhores oportunidades	better opportunities	[m'ʎɔɾzɔpɾtuni'dadʃ]
país onde	country where	['pɐiz'õd]

Table 4: *Examples of word sequences in Task 4.*

Word	Translation	Transcription
vinte à hora	twenty per hour	['vĩta'ɔɾɐ]
uma antiga colega	an old colleague	[ũmã'tigɐk ^{w}lɛgɐ]
sabe que o	knows that the	['sabkɹju]
penso que a	I think that the	['pẽskɹjɐ]
resultado do encontro	meeting result	[rzuɻ'tadːwẽ'kõtr^{w}]
sede de trabalhar	thirst to work	['sedːtɾɐbɐ'ʎaɾ]
tentar a sorte	try your luck	[tẽ'taɾɐ'sɔɾt]
mil euros	a thousand euros	['mil'ewɾɔʃ]

since they were also taken from exercises concluded in previous classes. Table 3 illustrates the 4 categories of our target processes: vowel encounters[19] (top row) and the production of /S/ as [ʃ], [ʒ] or [z] due to voicing assimilation [17] (second, third and fourth rows, respectively).

A second task performed in the same day was similar to the previous one, but targeted more complex connected processes. In this case, we tested haplologies (deletion of an equal or similar syllable as in "Cam[po] Pequeno"), occurrence of [J] and re-syllabification with liquids [ɾ] and [ɻ]. The selected news video, of 02:36 minutes, included real interviews which means students were exposed to casual speech [1]. We selected 11 target word sequences, totaling 29 different words. In this task, we tested stressed and unstressed words, ranging between 1 to 6 syllables. Table 4 illustrates examples of the 4 selected categories [19], [16]: vowel encounters (row 1), occurrence of [J] (row 2), haplologies (row 3), and resyllabification with liquids [ɾ] and [ɻ] (row 4). The experiment lasted around 10 minutes and was implemented two weeks after the previous one. The control group performed the same sequence of tasks in different moments. The experiments were applied in silent rooms such, as libraries and study or meeting rooms.

3.3. Production Data

Students' speech productions were collected from their final oral exam. In this test, students had to express their opinion about a picture or a topic taught in class. This exercise lasted 12 minutes. We analysed a small sample, in order to detect any occurrence of vowel reduction or any other *sandhi* phenomena. Our aim in analyzing spontaneous speech data was to further understand if the students would acquire and produce vowel reduction or connected speech processes at the end of the course. The data is still undergoing a more thorough process of analysis, which could not be fully performed for this paper.

Table 5: *Percentage of correct answers of each experiment for Mandarin speakers.*

Tasks	Correct Answers	Mean	S.D.
1 Single word w/out reduction	94%	9,4	0,8
2 Single word with reduction	65%	6,5	1,2
3 Simple connected speech	31%	2,5	1,7
4 Complex connected speech	16%	1,6	1,0

Table 6: *Percentage of correct answers of each experiment for the control group.*

Tasks	Correct Answers	Mean	S.D.
1 Single word w/out reduction	100%	10	0
2 Single word with reduction	100%	10	0
3 Simple connected speech	93%	7,4	0,8
4 Complex connected speech	98%	9,8	0,4

4. Results

4.1. Results of the perception experiments

Students' answers revealed they use several orthographic forms to transcribe the same word. We considered an answer correct when the word was orthographically correct or when the grapheme could correspond to the phonetic transcription of the intended grapheme, for instance: alternations between the grapheme "i" and "e" both corresponding to the sound [i] [23] and changes in the order of the segments to simplify consonant clusters [22]. The following tables show the percentage of correct answers calculated for each experiment for Mandarin speakers (Table 5) and the control group (Table6).

As shown in the Table 5, experiments with isolated word identification tasks had higher rates (94% and 65%) than experiments involving word identification tasks in continuous speech (31% and 16%), as expected [1], [8]. Task 1 revealed that students identified words correctly in a satisfactory way (94%). This means that the occurrence of all the phonetic segments was useful to listeners to recognize and identify most words. In task 2, the percentage decreased to 65%. Vowel reduction [17] is not a common process in Mandarin [14], [15]. So, this decrease revealed that vowel reduction within a word has a great impact on word identification process at this level.

The gap between tasks is even more evident in task 3 with only 31% of correct answers, showing that word identification is affected when vowel reduction, voicing assimilation and vowel encounters (common EP connected speech processes) occurred across word boundaries. Even though students knew the target words they did not recognize them in continuous speech [2].Thus, students might be using cues to segment speech that result in their L1 but not in EP [3], [8].

In task 4, we tested vowel encounters plus haplologies, insertion of [j] and re-syllabification with liquids [ɾ] and [ɻ] which means students dealt with stressed and unstressed words and with the increase of the number of target words. The rate of correct answers was 16%. This decrease shows that the occurrence of complex connected speech processes plus vowel reduction, always present within words and across word boundaries, heavily compromised word identification. In addition, students

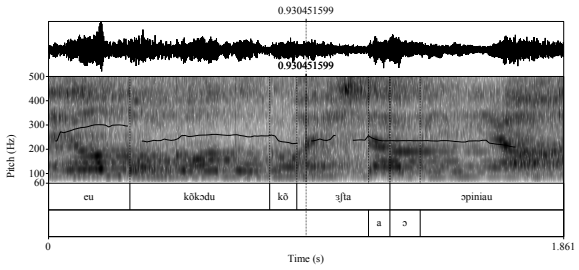


Figure 1: *Spectrogram of the sequence "eu concordo com esta opinião (I agree with this opinion).*

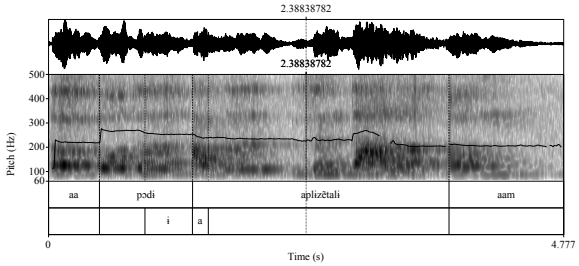


Figure 2: *Spectrogram of the sequence "pode apresentar amm" (could present um).*

were exposed to casual speech which is less careful [1] than the speech in a news report, as in task 3. As shown in the Table 6, control group had 100% for tasks 1 and 2. This highlights the fact natives do not have problems recognizing isolated words with or without vowel reduction. In addition, native participants rated 93% for task 3 and 98% for task 4, showing they did not have greater difficulties processing connected speech, in comparison with the test group. Nevertheless, these results reveal that some parsing problems may arise, mostly due to listening and writing at the same type.

4.2. Results of the production data

In order to understand the impact of all these phenomena in students' acquisition trajectory it was interesting not only to understand how word identification is affected, but also to know if students could produce vowel reduction and other phonetic processes in spontaneous speech.

Results show that students do not produce vowel reduction or any other *sandhi* process. Figure 1 and Figure 2 illustrate examples showing that vowel reduction is missing, for instance, the deletion of [u] ("concord[u]" in Figure 1) and [i] ("apr[i]sentar" in Figure 2) in unstressed positions. In addition, there is no vowel reduction in vowel encounters, for instance, the deletion of the final unstressed sound and the consequent articulation process with the first sound of the following word (Est[o]pinião, in Figure 1 and pod[ɐ]presentar in Figure 2), as expected in fluent EP [19]. Our results with spontaneous speech are consistent with the results described in[22], for reading tasks. However, in our study listening and segmenting words is heavily impacted by the presence of vowel reduction, the speech rate, and even the degree of vowel reduction and other *sandhi* phenomena. In a classroom context, the students of an intermediate level are not able to fully understand or produce several *sandhi* processes, mostly vowel reduction ones. This task is even more constrained due to spontaneous speech and didactic materials extracted from real contexts, an usual practice in classroom environments. It is interesting, however, to notice that some students apply re-syllabification with [ɾ] in fixed expressions such as "po[ɾ]exemplo" ("for instance") or "po[ɾ]isso" ("so that"). This means students learned that specific structure as a fixed lexical expression and they do not apply the process in other contexts. In the production data it is also worth mentioning the very distinctive prosodic and intonational patterns. The students, even the highest scored student, utter every stress word as a prosodic unit, with frequent pauses and stressing words which are unstressed, as if each word received a tone [20].

5. Conclusions

In this paper we investigated the impact of vowel reduction and other phonetic-phonological phenomena in Chinese L2 learners of Portuguese at the intermediate level (B1). The aim of this study was to understand the influence of these phenomena in word identification in continuous speech. From a didactic point of view, our study also aims to understand whether the acquisition of vowel reduction and connected speech processes takes place and how listening skills can be trained. The majority of EP L2 courses do not address these phonetic processes. Thus, we applied our study in a classroom setting with 12 native speakers of Mandarin and 12 EP native speakers, our control group. We designed two experiments concerning four word identification tasks involving these phenomena in increasing degrees of difficulty. Results show that the rate of correct answers for each task (in percentage) matched our initial expectations: decreasing from 94% to 65%, when single words were produced with vowel reduction, and substantially decreasing across word boundaries (31% and 16%). In contrast, the group of native speakers had higher rates: 100% in single word identification tasks similar rates in tasks testing connected speech (93% for task 3 and 98% for task 4). The results highlight that word recognition can be a challenge for L2 learners, due to the occurrence of vowel reduction (even within a word) and connected speech processes [8], [9], [10], [2]. Students are transferring properties from their L1 to their L2 and this is particularly true for lower proficiency levels [1], [22].

Production analysis reveals that learners do not produce vowel reduction nor any other phonetic-phonological phenomena across word boundaries, except within fixed expressions such as "por exemplo" (for instance), in which they apply re-syllabification with [ɾ]. The phonetic-phonological component in second language teaching should include some training concerning segmental and suprasegmental processes using authentic speech materials [9], [25]. The exposure to all these processes will improve their listening skills [8], [2]. The didactic outcomes of the experiments may contribute to the development of teaching materials for EP as L2, focusing on these processes.

Future work will tackle the comparison with other levels, in order to understand the degree of fluency when considering vowel reduction and other *sandhi* phenomena, always in a classroom environment and using the continuum from read-to-spontaneous speech.

6. Acknowledgements

This work was supported by national funds in Portugal through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020.

7. References

[1] A. Cutler, *Native Listening: Language Experience and the Recognition of Spoken Word*. Massachusetts: The MIT PRESS, 2012

[2] M. Reed, "Teaching Tip: Listening skills instruction: Practical tips for processing aural input," in *PPSLLTC 2019 – 10th Pronunciation in Second Language Learning and Teaching Conference, September, Ames, IA, Proceedings*, 2019, pp. 401–412.

[3] E. Altenberg, "The perception of word boundaries in a second language, " *Second Language Research - SECOND LANG RES*, vol. 21, no. 2, pp. 325–358, 2005.

[4] M. Broersma, "Perception of familiar contrasts in unfamiliar positions," *Journal of the Acoustical Society of America*, vol. 117, no. 2, pp. 380-390, 2005.

[5] J. Flege and c. Wang, "Native-language phonotatic constraints affect how well Chinese subjects perceive the world-final /t/-/d/ contrast," *Journal of Phonetics*, vol. 17, no. 2, pp. 299-315, 1989.

[6] C. Crowther and V. Mann, "Native language factors affecting use of vocalic cues to final consonant voicing," *Journal of the Acoustical Society of America*, vol. 92, no. 4, pp. 711–722, 1992.

[7] M. Boersma and A. Cutler, "Phantom word recognition in L2," *System*, vol. 36, no. 5, pp. 22–34, 2008.

[8] L. Vandergrift and C. Goh, *Teaching and Learning Second Language Listening: Metacognition in Action*. New York: Routledge, 2012.

[9] H. Mitterer and A. Tuinman, "The role of native-language knowledge in the perception of casual speech in a second language," *Frontiers in Psychology*, vol. 3, pp. 249, 2012.

[10] T. Snijders, V. Kooijman, A. Cutler and P. Hagoort, "Neurophysiological evidence of delayed segmentation in a foreign language," *Brain Research*, vol. 1178, no. 6, pp. 35–51, 2007.

[11] J. Field, "Revising segmentation hypotheses in first and second language listening," *System*, vol. 36, no. 7, pp. 35–51, 2008.

[12] A. Weber and A. Cutler, "First-language phonotatics in second-language listening," *Journal of the Acoustical society of America*, vol. 119, no. 8, pp. 597–607, 2006.

[13] L. Maastricht, *Second Language Prosody: Intonation and Rhythm in Production and Perception*. Tilburg University: Phd Thesis, 2018.

[14] S. Duanmu, *The Phonology of Standard Chinese*. Oxford: Oxford University Press, 2007.

[15] Y. Lin, *The Sounds of Chinese*. Cambridge: Cambridge University Press, 2007.

[16] M. Mateus, E. Andrade, *The Phonology of Portuguese*. Oxford: Oxford University Press, 2000.

[17] M. Mateus, I. Falé and M. Freitas, *Fonética e Fonologia do Português*. Lisboa: Universidade Aberta, 2005.

[18] S. Correia, S. Frota, J. Butler and M. Vigário, "Word stress perception in European Portuguese," in *INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association, August 25-29, Lyon, France, Proceedings*, 2013.

[19] E. Andrade and M. Viana, "Que horas são às (1)3 e 15?," *Actas do VIII Encontro Nacional da Associação Portuguesa de Linguística Lisboa: APL*, no. 8, pp. 59–66, 1994.

[20] W. Wang and K. Li, "Tone 3 in Pekinese," *Journal of Speech and Hearing Research*, no. 10, pp. 629–636, 1967.

[21] P. Mok, "On the syllable-timing of Cantonese and Beijing Mandarin," *Chinese Journal of Phonetics*, vol. 2, pp. 148–154, 2009.

[22] C. Zhou, M. Cruz and S. Frota, "O ritmo da interlíngua na produção do Português Europeu por falantes chineses," *Revista da Associação Portuguesa de Linguística*, no. 3, pp. 423–435, 2017.

[23] A. Castelo and R. Santos, "As vogais do português entre aprendentes chineses e suas implicações no desenvolvimento de um programa de português," *Português como língua estrangeira, de herança e materna: abordagens, contextos e práticas*, pp. 123–136, 2017.

[24] L. Grant, *Pronunciation Myths: Applying Second Language Research to Classroom Teaching*. Michigan: University of Michigan Press, 2014.

[25] A. Castelo, "Ensino da componente fonético-fonológica: uma síntese e um exemplo de português para estrangeiros," *Revista de Estudos Linguísticos da Universidade do Porto*, no. 12, pp. 41–71, 2017.

[26] P. Boersma and D. Weenink, "PRAAT: Doing phonetics by computer," *Glot International*, vol. 5, pp. 341–345, 2001.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Nativeness Assessment for Crowdsourced Speech Collections

Diogo Botelheiro[1,2], *Alberto Abad*[1], *Rui Correia*[2], *João Freitas*[2]

[1]INESC-ID/Instituto Superior Técnico, Universidade de Lisboa, Portugal
[2]DefinedCrowd Corporation

{diogo.botelheiro, correia,joao}@definedcrowd.com, alberto.abad@inesc-id.pt

Abstract

Access to large amounts of annotated data is a challenge for companies developing high-quality AI-based services. Crowdsourcing presents itself as a solution to this growing need for training data, by gathering and distributing work across a large pool of human contributors. This, however, comes at a cost: the difficulty to source the right crowd and maintain data quality. Regarding speech, a critical aspect of data quality relates to the verification of crowd participants as native speakers of a specific language. This work investigates the use of automatic Nativeness Classification (NC) solutions to tackle this problem, integrating a variant-sensitive nativeness classifier component in the speech collection pipeline for Portuguese (European and Brazilian variants) and English (American, British and Indian). By rating individual recordings according to their nativeness, it is possible to both automatically discard substandard work and prevent certain contributors to continue to participate in the collection. Herein, three different speaker-embedding-based frameworks are tested: i-vector, x-vector, and h-vector. Results show that the proposed system based on h-vector outperforms the baseline system with a 8% relative improvement.

Index Terms: nativeness classification, crowdsourcing, deep neural networks, x-vector

1. Introduction

Human-machine collaboration is moving towards more natural means of interaction, in particular via speech. Several speech-based commercial applications exist nowadays, ranging from personal assistants (Siri[1]), dictation systems (Otter.ai[2]), or home automation (Google Assistant[3]). Such technologies are based on high-performance Automatic Speech Recognizers (ASR) [1], which depend on large amounts of training data to improve and to have good performance in new markets (domains/languages). While traditionally such data resources were collected on-site, with experts, the massive amounts of data needed to support these systems deems this collection strategy unfeasible.

To respond to these needs, crowdsourcing emerged as a more scalable (both faster and less costly) approach to speech data collection [2]. Briefly, the crowdsourcing paradigm consists of making available a set of Human-Intelligence Tasks (HITs) to a large pool of contributors, typically via an online-platform. Upon successful completion, a reward or a payment is given to the contributors in an amount proportional to their participation. Given these particularities, crowdsourcing presents a new set of challenges, pertaining to both the loss of control by data requesters and the exploitation attempts by some contributors. More particularly, in the case of speech data collections, requirements such as recording/noise conditions, demographic balancing (age/gender), or nativeness, need to be ensured.

In the crowdsourcing field, these issues are commonly addressed by submitting each generated recording to further human validation [3]. In other words, a common speech data collection pipeline is composed of two steps:

- **Generation Step** - contributors are requested to read and record a given prompt;
- **Validation Step** - contributors are requested to validate certain aspects of a given audio (previously recorded by other contributor), such as the absence of background noise, and/or the speaker's nativeness degree.

However, this validation step increases the price and time to complete the collection. In the circumstances of the amounts of data necessary to train a state-of-the-art ASR model, this is a non-negligible cost since you need to validate thousands of hours of speech. Furthermore, if all the necessary components that need to be validated are included in one single step, this adds to the contributors' cognitive load. As a consequence, it requires higher payment for each of the individual tasks and increases the chances of errors in the validation.

In an effort to reduce the human validation load for the collection of speech data, this work addresses one of the most common validation components: nativeness. The hypothesis is that by integrating an automatic nativeness classifier, it is possible to either remove all the human validation concerning nativeness from the pipeline or provide hints of fraudulent behavior, reducing the amount of data that goes through human validation.

The work presented in this paper was done using real data from DefinedCrowd's[4] prorietary crowdsourcing platform – Neevo[5]. Due to the importance of generalizing language in speech collections, we explore two distinct languages and several variants of those languages, up to a total of five language-locale pairs: Portuguese (European and Brazilian variants) and English (American, British, and Indian variants).

The remainder of this paper is structured as follows: Section 2 sets some terminology and presents background work in the areas of Crowdsourcing and Nativeness Classification; Section 3 describes the experimental setup; Section 4 presents the results; and Section 5 concludes with a discussion of the results and some remarks about future work.

2. Background and state-of the-art

This section further describes the background in which this work was developed on (Section 2.1) and the state-of-the-art of the Nativeness Classification task (Section 2.2).

[1]https://www.apple.com/siri/
[2]https://www.otter.ai
[3]https://assistant.google.com/
[4]https://www.definedcrowd.com/
[5]https://www.neevo.ai/

10.21437/IberSPEECH.2021-5

2.1. Crowdsourcing

The introduction of the term crowdsourcing appears in 2006 by Jeff Howe [4] referring to the increasing practice of outsourcing tasks to the internet as a distributed procedure over various users. Crowdsourcing allows leveraging the so-called *wisdom of the crowds* [5]: the combined knowledge of potentially large groups of individuals. Common tasks approached with crowdsourcing are labeling images, translating or transcribing text, or recording speech, to name a few.

As previously mentioned, the current work was based on Neevo's crowdsourcing platform. From the contributors' point of view, participation is divided into four phases:

- **Registration** - contributors sign up, providing demographic (age/gender) and language data (including reading, writing, and speaking proficiency per language);
- **Work Selection** - depending on their qualifications, contributors see the matching tasks, which are organized into *Jobs*. A Job is a set of tasks (HITs) with a common goal, for instance, "Record yourself reading sentences in European Portuguese", or "Validate the English US recordings in terms of nativeness and noise". When a contributors accepts a Job they are referred to as *Job Members*;
- **Execution** - Job Members read the instructions of the Job and perform the HITs that are still available, usually in the order of the hundreds or thousands. An instance of a submitted HIT is called a *HIT Execution*;
- **Payment** - upon successful completion of the work (which can be dependent on subsequent validation jobs), the contributor is paid accordingly.

In the case of a Speech Collection, as also already mentioned, there are typically two jobs involved: a generation job and a validation one. The generation job can ask for spontaneous speech (where the contributor should talk about a topic for a certain amount of time) or for scripted speech (where the contributor reads a sentence). The validation job typically encompasses all the aspects that need to be validated (which can include text-audio match, background noise, and nativeness). In other words, if *any* of the aspects is not verified, the HIT Execution is canceled and should be re-recorded (not necessarily by the same contributor). Given the sensitivity of this decision, each HIT in the validation job is usually assigned to several distinct Job Members (two or three), in order to analyze for consistency and agreement between their answers.

2.2. Nativeness Classification

Nativeness classification is a subject that has been investigated for the past twenty five years. It is well-known that the presence of non-native speakers in the training set pose problems for speech recognition models, typically degrading their performance [6]. Recent literature on binary classification of nativeness uses distinct techniques. In studies like Shriberg et al. [7], the authors adress NC by applying effective speaker recognition methods based on Maximum Likelihood Linear Regression (MLLR), prosodic information, phone N-gram, and word N-gram features. Combining the different systems allowed to achieve an reasonable Equal Error Rate (EER) for detecting American English non-native speakers. Lopes et al. [8] developed a nativeness classifier using TED talks. A combination of acoustic and prosodic cues led to a good performance. In Mehrabani et al. [9], another implementation based on prosodic features, the authors were able to exceed the baseline accuracy of a Gaussian Supervector by over 10.0%. Another approach from Ribeiro et al. [10] developed several feature sets, including i-vectors, phonotactic models and n-grams counts based features. The results were superior from the presented baseline, with 44% improvements compared to results obtained by Honig et al [11]. Also introduced by Rajpal et al. [12] the use of longer duration cepstral features, namely Mel Frequency Cepstral Coefficients (MFCC) and auditory filterbank features learned from the database using Convolutional Restricted Boltzmann Machine (ConvRBM), allowed for accuracy improvements in the order of 40%.

3. Experimental Setup

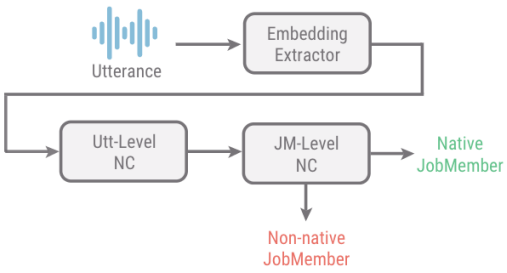


Figure 1: *Proposed Architecture.*

Figure 1 presents the architecture of the solution proposed to accomplish the goal of integrating nativeness assessment into a crowdsourcing speech data collection pipeline. The system is divided into three main components:

- **Embedding Extractor** - transforms the input speech data into a less-dimensional (vectorial) space;
- **Utterance-Level Nativeness Classification** - assigns a nativeness score to each HIT Execution;
- **Job Member-Level Nativeness Classification** - taking into account all the executions from any given Job Member, aggregates them providing a score on the likelihood of the Job Member being either a native or non-native speaker.

The following subsections will describe each component in detail, and introduce the experiments carried out for each of them. The work developed herein was based on the open-source toolkit Kaldi[6] for the feature extraction, data augmentation, and scoring, and TensorFlow[7], for neural network training.

3.1. Embedding Extractor

Advancements in the Automatic Language Identification field also impacted the Nativeness Classification field because of their similarities in intrinsic properties of each language [13, 14]. Nevertheless, with the emergence of Big Data [15], neural networks begun to generate interest. For today's Acoustic Models state-of-the-art, most systems use DNN for the embedding extractor [16], known as x-vector. Currently, to our knowledge, there is no published work with x-vector applications on NC that way, it would be interesting to apply the actual state-of-the-art of NLI [16, 17] to our research. The embedding extractor,

[6]https://github.com/kaldi-asr/kaldi
[7]https://github.com/sun-peach/x-vector-kaldi-tf

responsible for providing a vectorial representation of the input, is the variable component in this work experiments. Four different systems were developed:

- i-vector - previously referred to as state-of-the-art, we use this framework as a method of comparison [13];
- x-vector - TDNN implementation [16];
- h-vector - CNN implementation without attention mechanism [18];
- h-vector + attention - CNN implementation with attention mechanism [18].

The embedding structure follows the work of [16], using ReLU layers for the TDNN and Leaky-ReLU layers for the CNN. Given the specific context in which the solution will be used, and in face of available data, a dedicated corpus was built to train all four embedding systems (despite the various open-source corpus for tasks like spoken language recognition and native classification [19, 20, 21]).

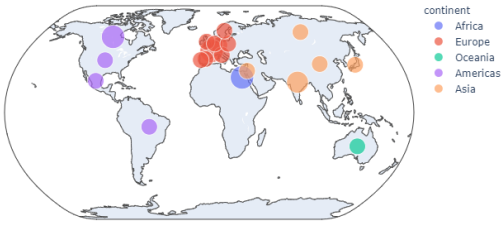


Figure 2: *Global representation for the embedding extractor corpus.*

To this end, validated data from different speech collections at DefinedCrowd was gathered, ultimately resulting in a dataset comprised of 25 language-locale pairs, with 132 hours of audio in total (113K utterances) represented in Figure 2. Data was balanced by trimming it to 5K utterances per language-locale pair (with the exception of ta_in, te_in, and zn_cn with only 1K utterances each). Data augmentation techniques were also used for robustness to noisy background, by applying additive noises using the Musan dataset [22]: 900 noises, 42 hours of music from various genres, and 60 hours of speech from twelve languages. This resulted in a final embedding training set of approximately 378K utterances.

In this work, we used 20-dimensional Mel-Frequency Cepstrum Coefficients (MFCC) with a frame-length of 25ms.

3.2. Utterance-Level Nativeness Classification

The second component that comprises the proposed solution is a nativeness classifier at the utterance level. In sum, it receives as input the vectorial representation of audio (embedding) and outputs a likelihood score with respect to its nativeness.

As mentioned in the introductory section, this work addresses five language-locale pairs: European Portuguese (pt-PT), Brazilian Portuguese (pt-BR), American English (en-US), British English (en-GB), and Indian English (en-IN). Consequently five corpora with nativeness information at the utterance level were built.

As in the case of embedding training, the data to train and test this component was extracted from real crowdsourcing data collections in the Neevo platform. Different strategies were applied for correctly labeling positive (native) and negative (non-native) labels. An utterance was considered positive when the

Table 1: *Division between train, dev and test dataset*

	Train set		**Dev set**		**Test set**	
Model	***#utt***	***#hours***	***#utt***	***#hours***	***#utt***	***#hours***
pt-PT	17300	31	2219	4	1167	2
pt-BR	15838	30	2792	5	1205	2
en-US	57522	112	4019	7	5098	8
en-GB	64495	111	5720	9	5028	8
en-IN	60997	113	5122	8	5790	5

job member producing them *a)* claimed (during sign-up) to be native of that language-locale pair, *b)* claimed (during sign-up) to live in the territory corresponding to the pair, and *c)* had **all** utterances approved in the validation job. On the other hand, utterances representing non-native speech were added through manual verification (by the author) based on work that was not accepted in the validation task. This manual process was necessary since, as already said, there are several reasons (other than nativeness) why a recording can be marked as invalid, including containing background noise, stuttering, hesitations, to name a few. Additionally, negative cases of each dataset were enriched with the positive cases of the other variants, i.e., for instance, utterances produced by pt-pt speakers were added to the non-native cases of the pt-BR set. Table 1 provides a description of the size of each dataset (after splitting into train, development and test).

For each target language, we extracted and computed the native and the non-native average vectors. Therefore, the representations are centered and projected using the training set. We start by applying the LDA with a dimension tuned to 200. After dimensionality reduction, the representations are length-normalized and modeled by PLDA, where we get two scores for each utterance. The score for native and non-native are normalized using adaptive s-norm [23]. Next, we performed a ratio between the scores, subtracting the native scores from the non-native to get the final score in Figure 3.

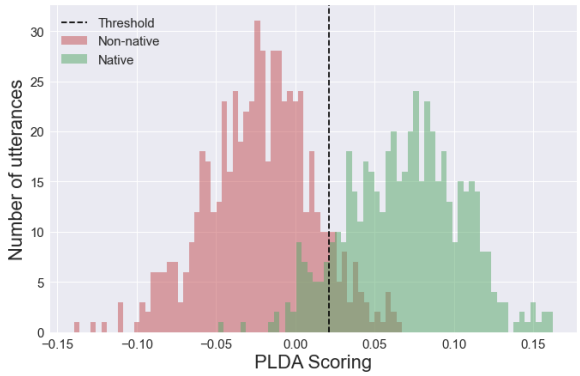


Figure 3: *Example of the result obtained for the h-vector framework for the European Portuguese language set.*

The optimal decision threshold corresponds to one which minimizes the Equal Error Rate (EER) on the dev set. The new utterances are processed by the scoring system and compared to the decision threshold. If the score is greater than the threshold, the utterance is classified as native and non-native otherwise.

Table 2: *Utterance-level classification results.*

	pt-pt		pt-br		en-us		en-gb		en-in	
System	**EER**	**F1-Score**	**EER**	**F1-Score**	**EER**	**F1-Score**	**EER**	**F1-Score**	**EER**	**F1-Score**
i-vector	0.15	0.76	0.32	0.69	0.16	0.83	0.18	0.81	0.16	0.84
x-vector (TDNN)	0.09	0.91	0.19	0.82	0.11	0.89	0.13	0.86	0.11	0.89
h-vector (CNN)	0.07	0.90	0.14	0.87	0.12	0.87	0.10	0.90	0.10	0.89
h-vector (CNN + attention)	0.02	0.96	0.11	0.90	0.10	0.89	0.09	0.91	0.09	0.90

3.3. JobMember-Level Nativeness Classification

Having a nativeness decision per utterance, the next step is to aggregate such information at Job Member level. A Job Member was considered to be a native speaker if the following two conditions were verified:

- Proportion of the Job Member's utterances considered to be native is above 50%;
- Average vector of all the Job Member's utterances is classified as native when using the designated language threshold.

If both conditions failed to be verified, the Job Member was classified as a non-native speaker. For the cases where only one condition was met, the job member was classified as 'ambiguous', potentially needing human verification.

4. Results

To better understand the performance of each component, results will be analyzed separately for utterance and Job Member levels.

Table 2 reports the results obtained from Utterance-Level NC. The first row shows the results from the NC state-of-the-art (i-vector framework), serving as a means of comparison. Results show that the x-vector framework outperforms the baseline in all languages under study, performing the best for the European Portuguese scenario (with an F1 of 0.91). However, its performance is still lacking for the use case in consideration, with 0.19 EER for the pt-BR set. One of the factors that may have hindered the performance was the lack of fine-tuning of some model parameters, such as the regularization coefficient. While the results of the h-vector framework alone are inconclusive (performing better than the x-vectors for some languages, and worse for others), the same is not true for its counterpart with attention mechanism. The h-vector formulation with self-attention mechanism surpassed both the UBM-GMM (i-vector, the state-of-the-art) and TDNN (x-vector) models, achieving a maximum F1 of 0.96 for the pt-PT set. It is also important to highlight the performance improvements achieved for the pt-BR data (0.32 vs. 0.11 EER and 0.69 vs. 0.90 F1).

To conclude on the performance of the Job Member-Level NC component, a system simulating the arrival of new utterances was set up. Job Member nativeness decisions were computed for different amounts of utterances, ranging from 1 utterance per Job Member, until the maximum available. The selection of which utterances to include in each step was random. The simulation was ran 1,000 times, and results per number of utterances were averaged. Figure 4 shows the results achieved. As in the utterance level component, the best results were observed for the pt-PT set, needing only three utterances to successfully classify all Job Members. American and British English achieved the next best performance, requiring eight utterances for accurate classification. Also as in the case of utterance classification, the worst performance was achieved for the Brazilian Portuguese data (eleven utterances needed). The results in this section take into account the test set comprising 45 Job Members for pt-PT, 59 for pt-BR, 121 for en-US, 100 for en-GB, and 96 for en-IN.

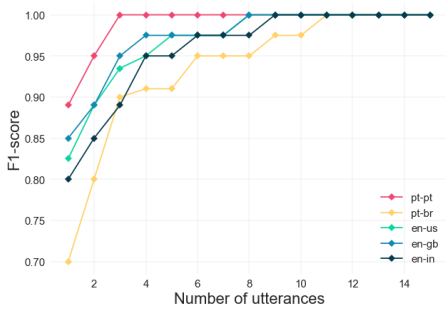


Figure 4: *Example of the simulation experiment for the h-vector (CNN + attention) framework.*

5. Conclusions

In this paper, we presented our work towards the deployment of a nativeness detection module into a crowdsourcing platform for quality control of speech collections. A corpus with more than 130 hours of audio with 25 languages was built to train four different embedding frameworks. We reported performance of Nativeness Classification on five different language-locale pairs along these four frameworks (utterance and Job Member-level). The h-vector solution with attention mechanism outperformed all approaches (including the i-vector baseline). Results show that the number of utterances needed to successfully verify Job Member nativeness vary across language, although having a ceiling of eleven utterances. In the field of crowdsourcing, this information can be used to optimize the speech collection pipeline, preventing substandard work from an early point, thus saving on both time and cost of the collection.

Future work includes the extension of the framework to a larger set of target languages and a deeper investigation of attention based extractor methods.

6. Acknowledgements

This work has been partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020.

7. References

[1] V. Zue, S. Seneff, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech communication*, vol. 9, no. 4, pp. 351–356, 1990.

[2] J. Freitas, J. Ribeiro, D. Baldewijns, S. Oliveira, and D. Braga,

"Machine learning powered data platform for high-quality speech and nlp workflows." in *INTERSPEECH*, 2018, pp. 1962–1963.

[3] V. Muntés-Mulero, P. Paladini, J. Manzoor, A. Gritti, J.-L. Larriba-Pey, and F. Mijnhardt, "Crowdsourcing for industrial problems," in *International Workshop on Citizen in Sensor Networks*. Springer, 2012, pp. 6–18.

[4] J. Howe, "The rise of crowdsourcing," *Wired magazine*, vol. 14, no. 6, pp. 1–4, 2006.

[5] J. Surowiecki, *The wisdom of crowds*. Anchor, 2005.

[6] L. Kilman, A. Zekveld, M. Hällgren, and J. Rönnberg, "The influence of non-native language proficiency on speech perception performance," *Frontiers in Psychology*, vol. 5, p. 651, 2014.

[7] E. Shriberg, L. Ferrer, S. S. Kajarekar, N. Scheffer, A. Stolcke, and M. Akbacak, "Detecting nonnative speech using speaker recognition approaches." in *Odyssey*, 2008, p. 26.

[8] J. Lopes, I. Trancoso, and A. Abad, "A nativeness classifier for TED talks," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5672–5675.

[9] M. Mehrabani, J. Tepperman, and E. Nava, "Nativeness classification with suprasegmental features on the accent group level," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[10] E. Ribeiro, J. Ferreira, J. Olcoz, A. Abad, H. Moniz, F. Batista, and I. Trancoso, "Combining multiple approaches to predict the degree of nativeness," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] F. Hönig, A. Batliner, and E. Nöth, "Automatic assessment of non-native prosody–annotation, modelling and evaluation," in *International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, 2012, pp. 21–30.

[12] A. Rajpal, T. B. Patel, H. B. Sailor, M. C. Madhavi, H. A. Patil, and H. Fujisaki, "Native language identification using spectral and source-based features." in *INTERSPEECH*, 2016, pp. 2383–2387.

[13] M. Senoussaoui, P. Cardinal, N. Dehak, and A. L. Koerich, "Native language detection using the i-vector framework." in *INTERSPEECH*, 2016, pp. 2398–2402.

[14] A. N. Uddin, M. A. Rahman, M. Islam, M. A. Haque *et al.*, "Native language identification using i-vector," *arXiv preprint arXiv:1811.05540*, 2018.

[15] G. George, M. R. Haas, and A. Pentland, "Big data and management," 2014.

[16] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors." in *Odyssey*, 2018, pp. 105–111.

[17] X. Miao, I. McLoughlin, and Y. Yan, "A new time-frequency attention mechanism for tdnn and cnn-lstm-tdnn, with application to language identification." in *INTERSPEECH*, 2019, pp. 4080–4084.

[18] Y. Shi, Q. Huang, and T. Hain, "H-vectors: Utterance-level speaker embedding using a hierarchical attention model," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7579–7583.

[19] J. Valk and T. Alumäe, "Voxlingua107: a dataset for spoken language recognition," *arXiv preprint arXiv:2011.12998*, 2020.

[20] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[22] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," 10 2015.

[23] D. E. Sturim and D. A. Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker verification," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–741.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Convolutional Recurrent Neural Networks for Speech Activity Detection in Naturalistic Audio from Apollo Missions

Pablo Gimeno, Dayana Ribas, Alfonso Ortega, Antonio Miguel, Eduardo Lleida

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain
{pablogj, dribas, ortega, amiguel, lleida}@unizar.es

Abstract

Speech Activity Detection (SAD) aims to correctly distinguish audio segments containing human speech. Several solutions have been successfully applied to the SAD task, with deep learning approaches being specially relevant nowadays. This paper describes a SAD solution based on Convolutional Recurrent Neural Networks (CRNN) presented as the ViVoLab submission to the 2020 Fearless steps challenge. The dataset used comes from the audio of Apollo space missions, presenting a challenging domain with strong degradation and several transmission noises. First, we explore the performance of 1D and 2D convolutional processing stages. Then we propose a novel architecture that executes the fusion of two convolutional feature maps by combining the information captured with 1D and 2D filters. Obtained results largely outperform the baseline provided by the organisation. They were able to achieve a detection cost function below 2% on the development set for all configurations. Best results were reported on the presented fusion architecture, with a DCF metric of 1.78% on the evaluation set and ranking fourth among all the participant teams in the challenge SAD task.

Index Terms: speech activity detection, convolutional recurrent neural networks, Fearless steps challenge, naturalistic audio

1. Introduction

Speech activity detection (SAD) aims to determine whether an audio signal contains speech or not, and its exact location in the signal. This constitutes an essential preprocessing step in several speech-related applications such as speech and speaker recognition, as well as speech enhancement. In many cases, the SAD is used as a preliminary block to separate the segments of the signal that contain speech from those that are only noise. This way, enabling the overall system to, for instance, performing speaker recognition only on speech segments.

A large number of approaches have been proposed for the SAD task. Starting with unsupervised approaches, some examples can be cited: based on energy [1], or based on the estimation of the signal long-term spectral divergence [2]. Traditionally, statistical approaches have been used with relevant results under the assumption of quasi-stationary noise. Several works rely on the extraction of specific acoustic features [3] [4]. Conversely, other methods are model-based [5] [6], aiming to estimate a statistical model for the noisy signal. Recently, deep learning approaches are becoming more and more relevant in the SAD task. The research presented in [7] implements a SAD system based on a multilayer perceptron with energy efficiency as the main concern. A deep neural network approach is used in [8] to perform SAD in a multi-room environment. In [9], new optimisation techniques based on the area under the ROC curve are explored in the framework of a deep learning SAD system.

Recurrent Neural Networks (RNN) are specially relevant in order to deal with temporal sequences of information. The Long Short Term Memory (LSTM) networks [10] are a kind of RNN that introduces the concept of the memory cell in order to learn, retain, and forget information in long dependencies. Some research has already proposed the use of LSTM networks to solve the SAD task. Authors in [11] presented an LSTM network to classify speech and non-speech segments in a noisy speech from Hollywood movies. A similar system is used in [12] to implement a noise-robust vowel based SAD. In this context, we have been able to obtain competitive results in the framework of diarisation tasks [13] [14] based on the properties of the Bidirectional LSTM (BiLSTM) classifier.

Convolutional Recurrent Neural Networks (CRNN) combine the capability of convolutional networks to capture frequency and time dependencies simultaneously seeking to extract discriminative features, and the capability of recurrent networks to deal with temporal series. Several examples of the use of CRNN in audio processing can be found in the literature [15] [16] [17]. Recently, CRNN have been proposed in the SAD task with relevant results. The approach presented in [18], based on the use of 2D convolutional layers, ranked first among all submissions in the 2019 Fearless steps challenge SAD task[1].

In this paper, we present our submission to the SAD task proposed for the Fearless steps challenge 2020. We introduce a supervised deep learning solution based on a CRNN that is fed with Mel filterbank energies as input. We explore alternatives for the convolutional layers, namely 1D and 2D filters. Then, we present a novel approach based on the fusion of two convolutional layers that combines the information of 1D and 2D filters to be processed by the RNN.

The remainder of the paper is organised as follows: a brief description of the Fearless steps challenge is given in section 2. Our CRNN based system proposal is described in section 3. The experimental setup for the challenge is introduced in section 4. Section 5 presents and discusses the results obtained. Finally, a summary and the conclusions are presented in section 6.

2. Fearless Steps challenge

The Fearless steps initiative has resulted in the digitisation of the original analog recordings from the Apollo space missions. Part of these data has been made available through the Fearless steps corpus, consisting of a cumulative 19,000 hours of conversational speech coming from the Apollo 11 mission [20]. Audio data belongs to 30 different communication channels, with multiple speakers in different locations. Most channels show a strong degradation with transmission noise or noise due to tape ageing. Furthermore, the signal-to-noise ratio (SNR) has a

[1] Results are no longer available online, but a summary of the best submissions can be found in [19]

10.21437/IberSPEECH.2021-6

strong variance, with levels ranging from 0 to 20 dB.

Aiming to motivate the research effort on this challenging domain, a series of annual challenges is being held proposing different speech related tasks. The inaugural Fearless steps challenge [19] took place in 2019, proposing the SAD task among other 4 different tasks. The focus on this first challenge was made on the development of unsupervised or semi-supervised systems. Only 20 hours of in-domain manually transcribed audio were available for the participants to use.

This new version of the challenge released in 2020 [21] changes its focus to the development of supervised systems, releasing around 80 hours of human labelled data through the training and development datasets. The SAD task is proposed again among other 5 different tasks. The fact that a larger amount of in-domain annotated data is available in this version opens a new possibility for supervised approaches such as the one proposed by this paper. Note that the use of out-of-domain data in these specific conditions, namely naturalistic audio and strongly degraded channels, could lead to poor results.

3. Proposed SAD system

3.1. Feature extraction

As input features for our proposed SAD system, we consider log Mel-filter bank energies. Namely, we use 64 log Mel-coefficients concatenated with the log energy of the frame. Note that as the input audio is sampled at $f_s = 8$ kHz, Mel filters span across the frequency range between 64 Hz and $f_s/2$. Features are computed every 10ms using a 25 ms Hamming window. As a final step, the mean and variance at feature level are used to normalise the corresponding file. All the alternatives developed to the SAD system proposal share the same set of features.

3.2. Neural architectures

In our submission to 2020 Fearless steps challenge for the SAD task we experimented with different neural architectures. In the following lines we briefly describe each of them.

As our baseline model, we choose a solution that is inspired by the SAD system proposed in our previous work in the diarisation framework [13]. It consists of an RNN block generated by stacking three BiLSTM layers with 128 neurons each. This block is then followed by a linear layer that generates the speech class score as a single neuron output.

The following architectures proposed are built on top of the RNN block from the baseline system, incorporating a set of convolutional layers working as a processing stage previous to the RNN block. The schematic representation of the proposed alternatives for the CRNN model is described in Figure 1. Note that the RNN block followed by a linear layer is shared by the three architectures. Then, the difference comes from the convolutional stage, that is implemented in three different ways:

- **Architecture A**: This model uses three 2D convolutional blocks processing the input features. Each of these blocks is integrated by a 2D convolutional layer with 3x3 or 5x5 kernel size and 64 filters. Then it is followed by a batch normalisation [22] and the application of a rectified linear unit (ReLU) [23] activation function. Finally, a max-pooling mechanism is applied considering a 4x1 stride, so that only the frequency axis is downsampled.
- **Architecture B**: This model similarly uses three 1D convolutional blocks. Even though, in this case, we experiment with different variations for the 1D convolutional layer. The first approach uses a kernel size of 3 in all the convolutions with no dilation. In the second implementation, each of the three layers uses kernel sizes of 5, 3 and 3 with dilations 1, 2 and 4 respectively. For the third approach, we experiment with the concept of group convolution, which has recently demonstrated its effectiveness in models such as ResNeXt [24]. This alternative employs a kernel size of 3 in all the convolutions but, in this case, these are implemented as 5 independent groups. This change does not affect the dimension of the input and the output feature maps but it reduces computational complexity and the number of model parameters. Finally, to obtain a comparable representation to the 2D setup, the convolutional layers have 256 output filters and a max-pooling mechanism is applied on the frequency axis using a 4x1 stride after the batch normalisation layer and a ReLU activation function.

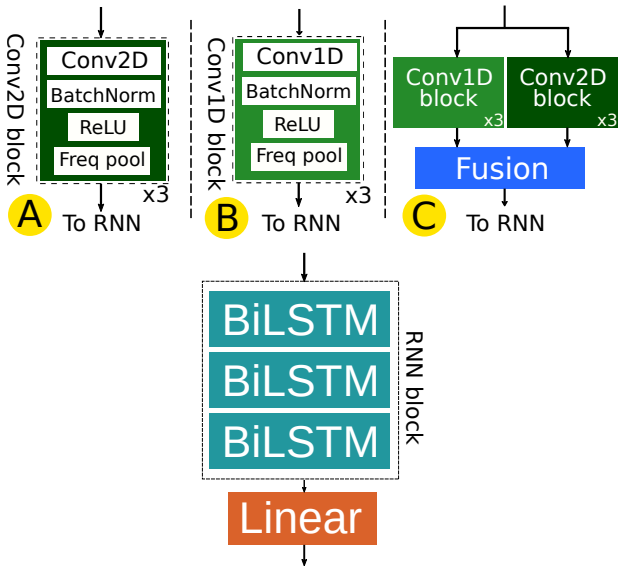


Figure 1: *Schematic representation of the different variations on the proposed convolutional recurrent neural network used for the SAD task.*

- **Architecture C**: Some previous work has already shown the combined capabilities of 1D and 2D convolutions applying system level fusion techniques [25]. This alternative proposes a novel approach to the CRNN model in the SAD task where we aim to combine the information extracted by two different convolutional branches. One consisting of three Conv2D blocks and other consisting of three Conv1D blocks, both implemented as described in the previous architectures. This fusion is done in an intermediate feature space, where both branches are then combined to be processed by the RNN block. The fusion block (depicted in blue) could be implemented in many different ways. In our experiments we test three different options: 1) a bilinear layer combining both convolutional branches, 2) the sum of the output of both branches, and 3) the concatenation of the output of both branches.

A common characteristic among all the models evaluated in this paper is that training and evaluation are performed using finite-length sequences. The input audio is separated in overlapping fragments of 3 second length and 2.5 second advance in order

to limit the delay of the dependencies that the network may take into account. The final prediction is generated by taking the first half of the overlapped part from the previous window, and the second half from the next window. This way the labels corresponding to the boundaries of each fragment are discarded as they may not be reliable. It must be noted too that, in all cases, the neural networks emit a SAD label for each frame processed at the input, one each 10ms in this case.

4. Experimental setup

4.1. Data description

The Fearless steps challenge follows open training conditions. Participants can use any available data in addition to the provided challenge data to train and tune their systems. However, in this work we have not used any additional datasets. Considering the specific domain of the audio, namely quite degraded channels and several kinds of transmission noises, we opted to use for training and development only the labelled data provided by the organisation. These data consist of 3 different partitions. In the following lines, we describe them and explain how they have been used in our submission:

- **Train**: Training subset is made of 125 files of around 30-minutes duration each. This makes a total of around 62.5 hours of audio. In our experiments we used 10% of these data for training validation. This way, all the proposed systems were trained with around 56 hours of audio from the train partition.
- **Development**: There are available 30 files of 30 minutes length for development purposes, resulting in around 15 hours of audio. This subset was used to obtain an empirical threshold, in order to minimise the detection cost function (DCF) metric. We also report our results on this subset.
- **Evaluation**: There are available 40 files of 30 minutes, which become 20 hours of audio for evaluation. We report our results on this subset as provided by the challenge organisation. The DCF metric obtained in the evaluation subset is the one used to rank the participants.

4.2. Training strategies

Models in this work are trained using Adam optimiser [26] with a learning rate that decays exponentially from 10^{-3} to 10^{-4} during the 20 epochs that data is presented to the neural network, with a minibatch size of 64. Cross entropy criterion is chosen as loss function, as usually done in classification tasks. Model selection is done choosing the best performing model in terms of frame classification accuracy using the validation subset. All the models in this paper have been developed using the PyTorch toolkit [27].

4.3. Evaluation metric

Two different errors can be considered when dealing with a SAD system: a false positive (FP), this is the identification of speech in a segment where the reference identifies non-speech, and a false negative (FN), this is the missed identification of speech in a segment where the reference identifies speech. With these two errors, we can define the probability of a false positive and the probability of a false negative according to the following equations:

$$P_{FP} = \frac{\mathrm{T}_{FP}}{\mathrm{T}_{\text{ref non-speech}}}\,, \qquad P_{FN} = \frac{\mathrm{T}_{FN}}{\mathrm{T}_{\text{ref speech}}}\,, \qquad (1, 2)$$

Table 1: *SAD results in terms of DCF metric on the development and evaluation partition, and number of trainable parameters for different systems considered for submission.*

System	# Param	DCF(%)	
		Dev	Eval
Organisation baseline [28]	-	12.50	13.60
RNN baseline	266K	2.02	2.54
A1 - CRNN 2D (3x3)	340K	1.65	2.07
A2 - CRNN 2D (5x5)	473K	1.67	2.28
B1 - CRNN 1D	421K	1.76	2.33
B2 - CRNN 1D dilation	455K	1.86	2.30
B3 - CRNN 1D groups	300K	1.76	2.46
C1 - CRNN fusion bilinear	641K	1.46	1.78
C2 - CRNN fusion sum	377K	1.60	1.89
C3 - CRNN fusion concat	411K	1.43	1.82

where T_{FP} and T_{FN} are, respectively, the total false positive time and total false negative time , $\mathrm{T}_{\text{ref non-speech}}$ represents the total annotated non-speech time in the reference, and $\mathrm{T}_{\text{ref speech}}$ represents the total annotated speech time in the reference.

In the SAD task of the Fearless steps challenge false negative errors are considered more important than false positive errors. This is shown in the primary evaluation metric for the challenge, the DCF, which is calculated as follows:

$$\mathrm{DCF}(\theta) = 0.75 P_{FN}(\theta) + 0.25 P_{FP}(\theta)\,, \qquad (3)$$

where P_{FN} is the probability for a false negative and P_{FP} is the probability for a false positive. Participants are responsible to choose a threshold (θ) that minimises the DCF.

5. Results

Table 1 presents the obtained results for the different systems submitted. We compare our RNN baseline system and the three proposed architectures to the baseline provided by the organisation [28]. Note that the organisation's baseline is based on a statistical approach. Concerning the fusion architectures, they use the best configurations achieved with the development set: A1 for the 2D setup, and B3 for the 1D setup, as it obtains similar results to B1 with a significantly smaller number of parameters. Results are reported in terms of DCF metric for both, development and evaluation partitions. To measure the level of complexity of the models, we present the number of trainable parameters for all submissions.

Regarding the results reported on the development partition, all our presented systems significantly outperform the baseline algorithm proposed by the organisation. Furthermore, all the systems that include a convolutional processing stage improve the performance compared to the RNN baseline. Our experimental findings are in line with the ones presented in [18], where 2D CRNN models provided better performance than 1D CRNN based SAD systems. In our case, using the 3x3 filter configuration we were able to obtain a DCF of 1.65%, which is better than all the 1D based systems evaluated. For the 1D convolution setup, it is interesting to mention the configuration using groups. A significant relative improvement of 12.77% compared to the RNN baseline is obtained being the CRNN model with the lowest number of parameters. These experimental results indicate that the combination of 1D and 2D feature maps is beneficial for our SAD system. Best overall results are obtained

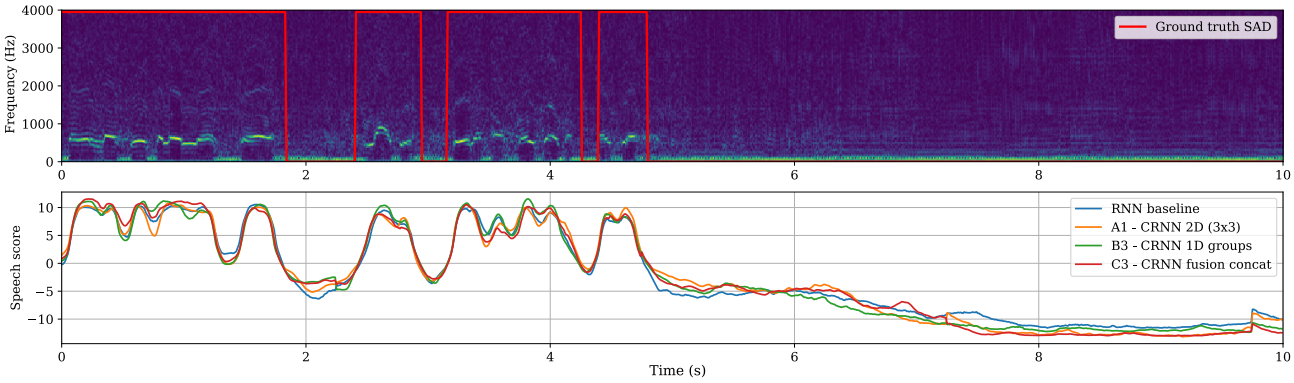


Figure 2: *Qualitative visualisation of SAD scores for the model alternatives described in the paper in a 10 seconds audio fragment extracted from file "FS02_dev_001". From top to bottom: audio spectrogram with the SAD ground truth overlapped in red colour and speech score for different SAD systems proposed.*

using the proposed fusion architecture. The most relevant result is the one that concatenates both convolutional outputs, obtaining the best result in the development set while keeping a lower number of parameters when compared to the fusion using a bilinear layer. With this setup we were able to achieve a competitive result for the development set with a 1.43% DCF metric.

Concerning the evaluation partition results, similar trends to the ones described in the development partition can be observed. In general terms, the behaviour for the three different kinds of architectures is consistent. Again, the 2D convolution setup outperforms its 1D equivalent. However, a difference can be observed in the 1D setup between development and evaluation results. While the groups configuration is the best performing in the development set, in the evaluation set this is done by the dilation setup. The fusion architectures show the best overall results, with a DCF metric below 2% for all the variations proposed. This solution allows our best submission to achieve a DCF metric of 1.78% using a bilinear layer as fusion method. This result was ranked in seventh position among the 28 challenge submissions to the SAD task, and fourth among the 7 participant teams[2]. Note that, unlike it was observed in the development set, the fusion based on concatenation offers a slight degradation in performance compared to the bilinear fusion, while keeping the number of parameters significantly smaller.

Additionally, Figure 2 presents a qualitative visualisation of the SAD performance for the best performing architectures in the development partition. It can be observed that, as it was expected, a high positive value in the neural network speech score is associated with a strong evidence of speech in the audio signal. In general terms, we can see that all the systems shown in Figure 2 can accurately capture the speech and non-speech segments in the audio fragment with the empirical threshold minimising the DCF being $\theta = -2$. Anyway, some inconsistencies can be observed between the ground truth and the speech scores on some points. This is probably due to labelling conditions, where a few non-speech segments in between two speech segments are labelled as speech. Proposed systems are able to capture this effect by showing a local minimum in the speech score for the mentioned fragments (see the first two seconds).

Focusing on the individual performance of the proposed systems, we can observe that the 2D and fusion systems show a lower score when a long fragment of non-speech is processed. On the other hand, the system based on 1D tends to output a higher score for speech fragments. In the case of transitions, no significant difference is observed among the systems presented, indicating a similar response between speech and non-speech fragments and vice-versa. It must be noted that all the systems presented in this paper achieve competitive results without introducing post-processing or smoothing techniques on the neural network output. As it can be observed from the depicted scores of Figure 2, the BiLSTM layers are able to impose a certain amount of inertia on the output so that the speech class score is smooth enough to be used by itself.

[2] https://fearless-steps.github.io/ChallengePhase2/Final.html

6. Conclusions

In this paper, we presented the ViVoLab submission to the SAD task of the Fearless Steps Challenge 2020. In this Challenge, we processed audio with degraded channels and several kinds of transmission noises from Apollo space missions. For our submission, we explored different CRNN models using 1D and 2D filters in the convolutional layers. We proposed a novel architecture that combines information coming from 1D and 2D filters in an intermediate feature space, which then is processed by the recurrent neural network. Obtained results largely outperform the baseline provided by the Challenge organisation. Our experimental achievements are in line with previous publications where 2D convolutions obtained better performance than equivalent 1D convolutions. Additionally, we showed that the combination of the information provided by 1D and 2D filters is beneficial for the SAD system, performing with the best results in the development and evaluation sets. Our best submission achieved a DCF metric of 1.46% and 1.78% respectively in the development and evaluation sets, ranking seventh among the 28 submissions to the challenge SAD task, and fourth among the 7 participant teams.

7. Acknowledgements

This work has been supported by the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the project TIN2017-85854-C4-1-R, Government of Aragón (Reference Group T36_20R) and co-financed with Feder 2014-2020 "Building Europe from Aragón".

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

8. References

[1] K.-H. Woo, T.-Y. Yang, K.-J. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[2] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[3] S. V. Gerven and F. Xie, "A comparative study of speech detection methods," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[4] J.-C. Junqua, B. Mak, and B. Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Transactions on speech and audio processing*, vol. 2, no. 3, pp. 406–412, 1994.

[5] J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 1965–1976, 2006.

[6] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Veselỳ, and P. Matějka, "Developing a speech activity detection system for the DARPA RATS program," in *Proc. Interspeech*, 2012, pp. 1969–1972.

[7] B. Liu, Z. Wang, S. Guo, H. Yu, Y. Gong, J. Yang, and L. Shi, "An energy-efficient voice activity detector using deep neural networks and approximate computing," *Microelectronics Journal*, vol. 87, pp. 12–21, 2019.

[8] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Deep neural networks for multi-room voice activity detection: Advancements and comparative evaluation," in *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016, pp. 3391–3398.

[9] Z.-C. Fan, Z. Bai, X.-L. Zhang, S. Rahardja, and J. Chen, "AUC optimization for deep learning based voice activity detection," in *Proc. IEEE ICASSP*, 2019, pp. 6760–6764.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies," in *Proc. IEEE ICASSP*, 2013, pp. 483–487.

[12] J. Kim, J. Kim, S. Lee, J. Park, and M. Hahn, "Vowel based voice activity detection with LSTM recurrent neural network," in *Proc. 8th International Conference on Signal Processing Systems*, 2016, pp. 134–137.

[13] I. Viñals, P. Gimeno, A. Ortega, A. Miguel, and E. Lleida, "Estimation of the number of speakers with variational bayesian PLDA in the DIHARD diarization challenge," in *Proc. Interspeech*, 2018, pp. 2803–2807.

[14] ——, "In-domain adaptation solutions for the RTVE 2018 diarization challenge," in *Proc. Iberspeech*, 2018, pp. 220–223.

[15] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE ICASSP*, 2015, pp. 4580–4584.

[16] F. Vesperini, L. Romeo, E. Principi, A. Monteriù, and S. Squartini, "Convolutional recurrent neural networks and acoustic data augmentation for snore detection," in *Neural Approaches to Dynamics of Signal Exchanges*. Springer, 2020, pp. 35–46.

[17] X. Huang, L. Qiao, W. Yu, J. Li, and Y. Ma, "End-to-end sequence labeling via convolutional recurrent neural network with a connectionist temporal classification layer," *International Journal of Computational Intelligence Systems*, pp. 341–351, 2020.

[18] A. Vafeiadis, E. Fanioudakis, I. Potamitis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, "Two-dimensional convolutional recurrent neural networks for speech activity detection," *Proc. Interspeech 2019*, pp. 2045–2049, 2019.

[19] J. H. Hansen, A. Joglekar, M. C. Shekhar, V. Kothapally, C. Yu, L. Kaushik, and A. Sangwan, "The 2019 Inaugural Fearless Steps Challenge: A Giant Leap for Naturalistic Audio," in *Proc. Interspeech*, 2019, pp. 1851–1855.

[20] J. H. Hansen, A. Sangwan, A. Joglekar, A. E. Bulut, L. Kaushik, and C. Yu, "Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon," in *Proc. Interspeech*, 2018, pp. 2758–2762.

[21] A. Joglekar, J. H. Hansen, M. C. Shekar, and A. Sangwan, "FEARLESS STEPS challenge (FS-2): Supervised learning with massive naturalistic apollo data," *Proc. Interspeech 2020*, pp. 2617–2621, 2020.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[23] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[24] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[25] H. Zeinali, L. Burget, and H. Cernocky, "Acoustic scene classification using fusion of attentive convolutional neural networks for DCASE2019 challenge," DCASE2019 Challenge, Tech. Rep., 2019.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 8024–8035.

[28] A. Ziaei, L. Kaushik, A. Sangwan, J. H. Hansen, and D. W. Oard, "Speech activity detection for NASA apollo space missions: Challenges and solutions," in *Proc. Interspeech*, 2014, pp. 1544–1548.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Dual-channel eKF-RTF framework for speech enhancement with DNN-based speech presence estimation

J. M. Martín-Doñas[1], *A. M. Peinado*[2], *I. López-Espejo*[3] *and A. M. Gomez*[2]

[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA),
Mikeletegi 57, 20009 Donostia/San Sebastian, Spain
[2]Dept. de Teoría de la Señal, Telemática y Comunicaciones, Universidad de Granada, Spain
[3]Dept. of Electronic Systems, Aalborg University, Denmark
jmmartin@vicomtech.org, {amp,amgg}@ugr.es, ivl@es.aau.dk

Abstract

This paper presents a dual-channel speech enhancement framework that effectively integrates deep neural network (DNN) mask estimators. Our framework follows a beamforming-plus-postfiltering approach intended for noise reduction on dual-microphone smartphones. An extended Kalman filter is used for the estimation of the relative acoustic channel between microphones, while the noise estimation is performed using a speech presence probability estimator. We propose the use of a DNN estimator to improve the prediction of the speech presence probabilities without making any assumption about the statistics of the signals. We evaluate and compare different dual-channel features to improve the accuracy of this estimator, including the power and phase difference between the speech signals at the two microphones. The proposed integrated scheme is evaluated in different reverberant and noisy environments when the smartphone is used in both close- and far-talk positions. The experimental results show that our approach achieves significant improvements in terms of speech quality, intelligibility, and distortion when compared to other approaches based only on statistical signal processing.

Index Terms: Dual-microphone smartphone, beamforming, extended Kalman filter, speech presence probability, deep neural network

1. Introduction

Speech-related services are ubiquitously available thanks to mobile devices such as smartphones. These devices are frequently used in reverberant and noisy environments, both in close-talk (CT) conditions (i.e., the smartphone is placed at the ear of the user) and far-talk (FT) conditions (i.e., the user holds the device at a distance from her/his face). This makes speech enhancement algorithms particularly necessary to improve speech quality and intelligibility on these challenging scenarios.

Current smartphones often embed several microphones. Particularly, a widely used layout consists of a primary microphone at the bottom and a secondary one at the top or back of the device. While beamforming techniques (i.e., spatial filtering) [1] can be used in these devices, the reduced number of microphones and their location limit the speech enhancement performance [2]. In these circumstances, postfiltering techniques can be incorporated to these devices [3, 4] to improve the noise reduction. Alternatively, other approaches employ single-channel filters exploiting dual-channel information. For example, the power level difference between channels was exploited in [5, 6] for noise estimation and reduction in CT conditions. On the other hand, in [7, 8] the coherence properties of the noise field were considered to estimate the noise statistics, needed by a Wiener filter, in FT conditions. In addition to speech enhancement, the dual-channel information has been exploited for other related speech processing tasks from a classical signal processing perspective, as feature enhancement in automatic speech recognition (ASR) systems [9, 10] and noise estimation [11]. Finally, the use of deep neural networks (DNNs) has also been explored on dual-microphone smartphones. For example, in [12, 13] a DNN-based feature enhancement approach was investigated in the context of noise-robust ASR for smartphones. On the other hand, in [14] we proposed a dual-channel DNN-based speech enhancement algorithm based on spectral mapping. Recently, this idea was evaluated in [15] for spectral masking using phase-sensitive masks and dual-channel features.

In previous works [16, 17, 18], we proposed a dual-channel speech enhancement framework, intended for smartphones, based on a beamforming-plus-postfiltering scheme. The main contribution of our approach was the estimation of the acoustic response between microphones using an extended Kalman filter (eKF) framework, which allows us to track these acoustic channels in reverberant environments. Moreover, noise estimation was performed using a speech presence probability (SPP)-based approach to update the noise statistics when speech was absent. This SPP estimation was carried out using statistical spatial models with a priori SPP information obtained from dual-channel information. On the contrary, in this work we propose the integration of DNN-based mask estimators [19, 20, 21, 22] for this task. The DNN, which is fed with dual-channel features based on power and phase differences, aims to improve the accuracy of the prediction. Our proposal is then evaluated on a dual-microphone smartphone under several noisy acoustic environments in CT and FT conditions. The results show that our approach achieves improvements in terms of speech quality, intelligibility, and distortion in comparison with other state-of-the-art approaches for dual-channel smartphones.

The remainder of this paper is organized as follows. In Section 2, we briefly review our dual-channel eKF-based framework for smartphones. Section 3 describes our DNN-based

This work has been supported by the Spanish Ministry of Science and Innovation Project No. PID2019-104206GB-I00/AEI/10.13039/501100011033 and the Spanish Ministry of Universities through the National Program FPU (grant reference FPU15/04161).

10.21437/IberSPEECH.2021-7

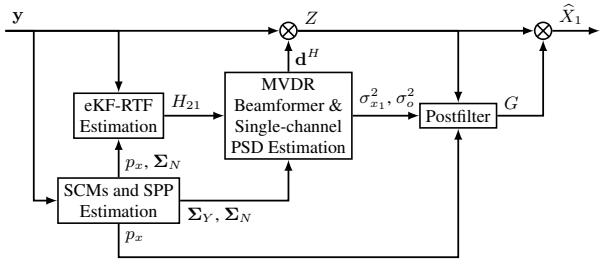


Figure 1: *Overview of the dual-channel speech enhancement algorithm for dual-microphone smartphones.*

mask estimator for SPP prediction and the dual-channel features evaluated. Then, in Section 4, the experimental framework and results are presented and analyzed. Finally, conclusions are summarized in Section 5.

2. Dual-channel speech enhancement based on an eKF-RTF framework

Let us consider the following multichannel observation model, in the short-time Fourier transform (STFT) domain, for the noisy speech signal acquired by a dual-microphone smartphone,

$$\mathbf{y}(t,f) = \mathbf{h}(t,f)X_1(t,f) + \mathbf{n}(t,f), \tag{1}$$

where $\mathbf{y}(t,f) = \begin{bmatrix} Y_1(t,f) & Y_2(t,f) \end{bmatrix}^\top$ and $\mathbf{n}(t,f) = \begin{bmatrix} N_1(t,f) & N_2(t,f) \end{bmatrix}^\top$ are the noisy speech and noise multichannel vectors (subscripts identify the primary and secondary microphones in the array), respectively, $X_1(t,f)$ is the clean speech signal at the primary microphone, $\mathbf{h}(t,f) = \begin{bmatrix} 1 & H_{21}(t,f) \end{bmatrix}^\top$ is the relative transfer function (RTF) vector, and t and f are the time frame and frequency indices, respectively. From now on, when possible, we will omit indices t and f for the sake of simplicity.

Our goal is to estimate the clean speech signal at the reference microphone, X_1, from the noisy speech observations. To do this, we apply our extended Kalman filter (eKF) dual-channel framework [18]. A diagram of the algorithm pipeline is depicted in Figure 1. The noisy speech signal is processed using a beamforming algorithm for noise reduction, as in $Z = \mathbf{d}^H\mathbf{y}$, where $\{\cdot\}^H$ stands for the Hermitian transpose operator. We use the minimum-variance distortionless response (MVDR) beamformer [1],

$$\mathbf{d} = \frac{\mathbf{\Sigma}_N^{-1}\mathbf{h}}{\mathbf{h}^H\mathbf{\Sigma}_N^{-1}\mathbf{h}}, \tag{2}$$

where $\mathbf{\Sigma}_N = E\left\{\mathbf{n}\mathbf{n}^H\right\}$ is the noise spatial covariance matrix (SCM), with $E\{\cdot\}$ representing the expectation operator over a random variable. At the beamformer output, signal Z represents the clean speech signal X_1 plus a residual noise with a power spectral density (PSD) given by $\sigma_o^2 = \left(\mathbf{h}^H\mathbf{\Sigma}_N^{-1}\mathbf{h}\right)^{-1}$.

As can be seen, MVDR needs estimates for H_{21} and $\mathbf{\Sigma}_N$. For RTF estimation, the already proposed eKF-RTF algorithm is applied [16, 18]. We first define $\mathbf{H}_{21}$ and $\mathbf{Y}_2$ as vectors that stack the real and imaginary components of H_{21} and Y_2, respectively. Then, the RTF vector is estimated using the following recursion,

$$\widehat{\mathbf{H}}_{21}(t) = \widehat{\mathbf{H}}_{21}(t-1) + \mathbf{K}(t)\left(\mathbf{Y}_2(t) - \boldsymbol{\mu}_Y(t)\right), \tag{3}$$

where $\mathbf{K}$ is the Kalman gain matrix and $\boldsymbol{\mu}_Y = E\{\mathbf{Y}_2\}$ is the expectation over the noisy speech at the secondary microphone. A detailed derivation of these terms can be found in [18]. For the noise statistics, we use a recursive estimator based on the speech presence probability (SPP) [23],

$$\widehat{\mathbf{\Sigma}}_N(t) = \alpha(t)\widehat{\mathbf{\Sigma}}_N(t-1) + (1-\alpha(t))\,\mathbf{y}(t)\mathbf{y}^H(t), \tag{4}$$

where $\alpha = \widetilde{\alpha} + (1-\widetilde{\alpha})\,p_x$ is an updating parameter that depends on the a posteriori SPP p_x, which ranges from 0 to 1, and $\widetilde{\alpha} = 0.9$ is a constant factor. Thus, the noise SCM is updated with the current noisy observation when speech is absent, while the previous value is kept when speech is present.

The speech signal Z is further processed using a postfilter which provides $\widehat{X}_1 = GZ$, where G is a single-channel gain function. In our proposal, we use the optimally-modified log spectral amplitude (OMLSA) estimator [24], which is defined as

$$G = (G_x)^{p_x}(G_n)^{1-p_x}, \tag{5}$$

in which G_n is the speech absence gain, set to -25 dB, and

$$G_x = \frac{\xi}{1+\xi}\exp\left(\frac{1}{2}\int_{\frac{\xi}{1+\xi}\gamma}^{\infty}\frac{e^{-u}}{u}du\right) \tag{6}$$

is the speech presence gain, with $\gamma = |Z|^2/\sigma_o^2$ being the a posteriori signal-to-noise ratio (SNR), $\xi = \sigma_{x_1}^2/\sigma_o^2$ the a priori SNR, and $\sigma_{x_1}^2$ the clean speech PSD at the primary microphone. This last PSD can be obtained using a maximum-likelihood estimator at the beamformer output [25],

$$\widehat{\sigma}_{x_1}^2 = \mathbf{d}^H\left(\widehat{\mathbf{\Sigma}}_Y - \widehat{\mathbf{\Sigma}}_N\right)\mathbf{d}, \tag{7}$$

where

$$\widehat{\mathbf{\Sigma}}_Y(t) = \tilde{\alpha}\widehat{\mathbf{\Sigma}}_Y(t-1) + (1-\tilde{\alpha})\,\mathbf{y}(t)\mathbf{y}^H(t) \tag{8}$$

is a recursive estimator of the noisy speech SCM. Finally, the gain function G is further processed by a musical noise reduction algorithm, as that described in [26], before applying it to the beamformed speech signal Z.

3. DNN-based a posteriori SPP estimation

As can be observed, the a posteriori SPP p_x plays a crucial role in our eKF-RTF framework. Not only that it controls the noise SCM estimation, but also the postfiltering proper performance depends on accurate SPP estimates. Besides, the RTF updating in (3) is only performed when speech presence is detected [18]. In our previous work [18], the a posteriori SPP was obtained using statistical spatial models that combine the use of multivariate Gaussian likelihoods (formulated for the noisy speech and noise signals) with a dual-channel a priori SPP estimator. The main drawback of this method lies on the assumptions made about the statistics of the signals, which can be inappropriate in realistic non-stationary environments.

In this work, we explore the use of DNN-based mask estimators [19] to directly compute the a posteriori SPP. In particular, we consider a convolutional recurrent network (CRN) [15] for the estimation of p_x. A diagram of the applied CRN architecture is depicted in Figure 2. As can be observed, the model comprises an encoder with five convolutional layers, a decoder with five deconvolutional layers, and an intermediate long short-term memory (LSTM) network. We use exponential linear units (ELUs) as non-linear functions in all the convolutional and deconvolutional layers except for the output layer, which uses the sigmoid function. A dropout layer is placed before the input to

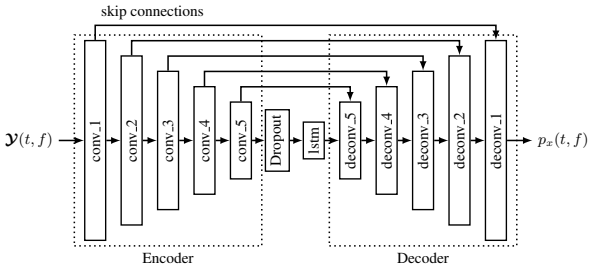


Figure 2: *Diagram of the CRN architecture used for the estimation of speech presence probability masks.*

the LSTM layer to help prevent overfitting. The convolutional and deconvolutional layers operate along the frequency dimension only, while the LSTM layer exploits the temporal dimension. Furthermore, we use skip connections that concatenate the output of each encoder layer to the input of each decoder layer. The CRN is trained using ideal binary masks from the reference channel as target features, using binary cross-entropy as loss function.

As network input features $\boldsymbol{\mathcal{Y}}(t,f)$, a set of different features exploiting spectral or spatial properties was considered in this paper. The main features of this set are the log-magnitude spectrum (LMS) of the primary channel, $\mathcal{Y}_{\text{LMS}}(t,f) = \log|Y_1(t,f)|$. An online normalization is applied to them using a time-recursive mean computation and subtraction at each frequency bin [27]. We also include additional features that make use of the inter-channel properties of the signals. In particular, we consider the spectral relation between the channels by using instantaneous power level difference (PLD) features, which are defined as

$$\mathcal{Y}_{\text{PLD}}(t,f) = \frac{|Y_1(t,f)|^2 - |Y_2(t,f)|^2}{|Y_1(t,f)|^2 + |Y_2(t,f)|^2}. \tag{9}$$

In addition, spatial properties of the signals are exploited by using inter-channel phase difference (IPD) features [28],

$$\boldsymbol{\mathcal{Y}}_{\text{IPD}}(t,f) = \begin{bmatrix} \cos\left(\theta_{y_1}(t,f) - \theta_{y_2}(t,f)\right) \\ \sin\left(\theta_{y_1}(t,f) - \theta_{y_2}(t,f)\right) \end{bmatrix}, \tag{10}$$

where θ_{y_1} and θ_{y_2} are the phases of the noisy speech signals at the reference and secondary microphones, respectively.

4. Experimental results

The TIMIT-2C-CT/FT database [18] was used to evaluate the proposed dual-channel algorithm. This database includes simulated dual-channel noisy speech recordings at 16 kHz acquired with a dual-microphone smartphone in both CT and FT conditions. Each condition (CT or FT) comprises three sets, i.e., training, validation, and test, with a total of 4800, 1200, and 2400 noisy speech samples, respectively. To simulate the noisy samples, clean speech signals from the TIMIT database [29, 30] are convolved with dual-channel acoustic responses obtained from a smartphone at different reverberant scenarios [16]. Then, the reverberated speech signals are mixed with noises at six SNRs from -5 dB to 20 dB (5 dB steps). For the training and validation sets, four reverberant and noisy environments are considered: car, bus, babble, and mall. For the test set, apart from the previous noise types, four additional environments are also evaluated: street, pedestrian street, bus station, and cafe. In addition, the reverberation level of the acoustic responses depends on the noisy environment.

Table 1: *Architecture of the CRN mask estimator. The feature size is indicated in the form feature maps × frames × frequency channels, being N_{in} the number of input features. Hyperparameters refer to kernel size, stride and output channels. For the LSTM layer, the number of hidden units is indicated.*

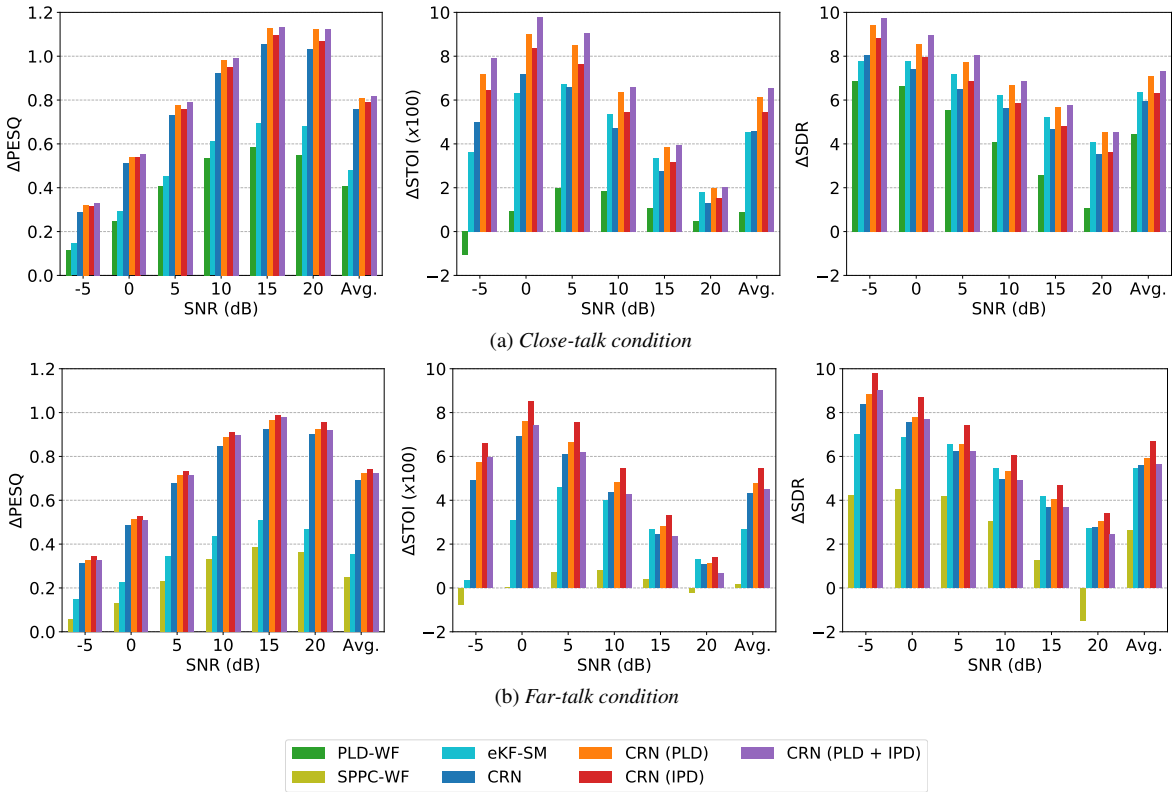
Layer name	Input size	Hyperparameters	Output size
conv_1	$N_{\text{in}} \times \text{T} \times 257$	1×3, (1, 2), 8	$8 \times \text{T} \times 128$
conv_2	$8 \times \text{T} \times 128$	1×3, (1, 2), 8	$8 \times \text{T} \times 64$
conv_3	$8 \times \text{T} \times 64$	1×3, (1, 2), 16	$16 \times \text{T} \times 32$
conv_4	$16 \times \text{T} \times 32$	1×3, (1, 2), 32	$32 \times \text{T} \times 16$
conv_5	$32 \times \text{T} \times 16$	1×3, (1, 2), 64	$64 \times \text{T} \times 8$
reshape_1	$64 \times \text{T} \times 8$	-	$\text{T} \times 512$
lstm	$\text{T} \times 512$	512	$\text{T} \times 512$
reshape_2	$\text{T} \times 512$	-	$64 \times \text{T} \times 8$
deconv_5	$128 \times \text{T} \times 8$	1×3, (1, 2), 32	$32 \times \text{T} \times 16$
deconv_4	$64 \times \text{T} \times 16$	1×3, (1, 2), 16	$16 \times \text{T} \times 32$
deconv_3	$32 \times \text{T} \times 32$	1×3, (1, 2), 8	$8 \times \text{T} \times 64$
deconv_2	$16 \times \text{T} \times 64$	1×3, (1, 2), 8	$8 \times \text{T} \times 128$
deconv_1	$16 \times \text{T} \times 128$	1×3, (1, 2), 1	$1 \times \text{T} \times 257$

Table 2: *Objective metric results for the noisy speech signals of the test set in the TIMIT-2C-CT/FT database, broken down by SNR and device use mode (CT or FT).*

Metric	Mode	SNR (dB)						Avg.
		-5	0	5	10	15	20	
PESQ	CT	1.09	1.11	1.23	1.45	1.81	2.27	1.49
	FT	1.07	1.11	1.25	1.50	1.88	2.38	1.53
STOI	CT	0.51	0.63	0.74	0.84	0.91	0.95	0.76
	FT	0.50	0.61	0.73	0.83	0.90	0.95	0.75
SDR	CT	-5.80	-0.81	4.19	9.15	14.02	18.70	6.58
	FT	-5.79	-0.80	4.19	9.15	14.03	18.70	6.58

For STFT computation, a 512-point DFT was applied using a 32 ms square-root Hann window with 50% overlap. The eKF-RTF framework implemented is the same as in [18]. The CRN network architecture used in our experiments is concisely described in Table 1. The ADAM optimizer [31] was used to train the DNN model. We used a batch size of 10 utterances, which were zero-padded to have the same number of frames. The dropout rate was set to 0.5 deactivation probability. Besides, the early-stopping procedure [32] was applied with a patience of 20 epochs.

The enhanced signal provided by our proposal is evaluated in terms of the following objective quality metrics: perceptual evaluation of the speech quality (PESQ) [33], short-time objective intelligibility (STOI) [34] and scale-invariant signal-to-distortion ratio (SDR) [35]. As a reference, Table 2 shows the results obtained in terms of these metrics when evaluating the noisy speech signals from the test set without any enhancement algorithm. For the CRN-based mask estimator, different combinations of input features were tested: using only LMS features (CRN), jointly integrating either PLD features (PLD) or IPD features (IPD), and fully integrating all the features (PLD+IPD). For comparison purposes, we also evaluated our framework with SPP estimation based on statistical models (eKF-SM) [18], and two single-channel Wiener filters relying on dual-channel information: the PLD-based filter (PLD-WF) [5] for the CT condition, and the SPP- and coherence-based filter (SPPC-WF) [7] for the FT condition. The results achieved by the tested methods (improvements obtained over the noisy speech results) are shown in Figure 3.



(a) *Close-talk condition*

(b) *Far-talk condition*

Figure 3: *PESQ, STOI and SDR differential results from the evaluation of the CRN-based SPP mask estimator with the different input features. The OMLSA postfilter with MVDR-based PSD estimation, and state-of-the art single-channel filters for smartphones, are also shown for comparison purposes. The plots show the increments obtained on the metrics with respect to noisy speech (see Table 2).*

In CT conditions, and according to Figure 3a, the CRN approach outperforms PLD-WF and eKF-SM, especially in terms of PESQ. Moreover, the CRN estimator benefits from the use of dual-channel features. PLD+IPD obtains the best results in terms of all the considered metrics. Between the dual-channel features considered, the CRN estimator mainly benefits from the PLD features, achieving similar results to those from PLD+IPD. This shows that the power difference between microphones can be a good indicator of speech presence in CT conditions. Although the CRN with IPD features improves with respect to the CRN approach, PLD+IPD slightly improves the variant including PLD features in STOI and SDR. Therefore, PLD features poses a good trade-off between performance and network complexity in CT conditions.

In FT conditions (Figure 3b), the CRN approach also achieves better results than the other evaluated methods. In this case, the variant including IPD features stands as the best choice, especially in terms of STOI and SDR. On the other hand, the utilization of PLD features slightly improves the CRN approach. Unlike in the CT scenario, PLD features seem not to provide enough information in FT conditions, as they tend to zero. Then, the phase difference is the main information source that allows to distinguish between speech and noise components. Finally, the combination of both types of dual-channel features does not yield improvements in comparison with using standalone features, either PLD or IPD ones. This can be explained by our CRN estimator not being able to deal with multiple input features in this challenging scenario. In particular, the use of additional PLD features may mislead the network, as they do not provide accurate information in this case. Thus, the IPD features stand as the best alternative.

5. Conclusions

In this paper, we have proposed a DNN-based SPP estimator that is integrated into our dual-channel eKF-RTF framework for speech enhancement on smartphones. Our approach allows for a more accurate prediction of the SPP probability thanks to the modeling capabilities of the DNN models and the use of dual-channel information. We use a convolutional recurrent neural network to exploit the spectral, spatial, and temporal properties of the speech signal. Two different dual-channel features were considered and tested: the instantaneous power level difference and the phase difference between channels. The proposed integrated scheme was compared with the same framework but using statistical spatial models for SPP prediction, as well as with other dual-channel speech enhancement algorithms from the state-of-the-art. The results show that the DNN-based mask estimator outperforms the rest of the evaluated approaches in terms of objective quality and intelligibility metrics. Among the considered spatial features, the PLD features show better performance in CT conditions, while the IPD features are more useful in FT conditions.

6. References

[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio,*

Speech and Language Processing, vol. 25, no. 4, pp. 692–730, 2017.

[2] I. Tashev, S. Mihov, T. Gleghorn, and A. Acero, "Sound capture system and spatial filter for small devices," in *Proc. InterSpeech*, 2008, pp. 435–438.

[3] E. Habets, S. Gannot, and I. Cohen, "Dual-microphone speech dereverberation in a noisy environment," in *Proc. IEEE International Symposium on Signal Processing and Information Technology*, 2006, pp. 651–655.

[4] C. Zheng, H. Liu, R. Peng, and X. Li, "A statistical analysis of two-channel post-filter estimators in isotropic noise fields," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 2, pp. 336–342, 2013.

[5] M. Jeub, C. Herglotz, C. Nelke, C. Beaugeant, and P. Vary, "Noise reduction for dual-microphone mobile phones exploiting power level differences," in *Proc. ICASSP*, 2012, pp. 1693–1696.

[6] V. B. Truong, D. M. Nguyen, and Q. H. Dang, "An MC-SPP approach for noise reduction in dual microphone case with power level difference," in *Proc. International Conference on Advanced Technologies for Communications*, 2014, pp. 292–297.

[7] C. M. Nelke, C. Beaugeant, and P. Vary, "Dual microphone noise PSD estimation for mobile phones in hands-free position exploiting the coherence and speech presence probability," in *Proc. ICASSP*, 2013, pp. 7279–7283.

[8] W. Jin, M. J. Taghizadeh, K. Chen, and W. Xiao, "Multi-channel noise reduction for hands-free voice communication on mobile phones," in *Proc. ICASSP*, 2017, pp. 506–510.

[9] I. López-Espejo, A. M. Gomez, J. A. González, and A. M. Peinado, "Feature enhancement for robust speech recognition on smartphones with dual-microphone," in *Proc. EUSIPCO*, 2014, pp. 21–25.

[10] I. López-Espejo, A. M. Peinado, A. M. Gomez, and J. A. González, "Dual-channel spectral weighting for robust speech recognition in mobile devices," *Digital Signal Processing*, vol. 75, pp. 13–24, 2018.

[11] I. López-Espejo, J. M. Martín-Doñas, A. M. Gomez, and A. M. Peinado, "Unscented transform-based dual-channel noise estimation: Application to speech enhancement on smartphones," in *Proc. IEEE Telecommunications and Signal Processing*, 2018, pp. 88–91.

[12] I. López-Espejo, J. A. González, Á. M. Gómez, and A. M. Peinado, "A deep neural network approach for missing-data mask estimation on dual-microphone smartphones: Application to noise-robust speech recognition," in *Advances in Speech and Language Technologies for Iberian Languages*, 2014, pp. 119–128.

[13] I. López-Espejo, A. M. Peinado, A. M. Gomez, and J. M. Martín-Doñas, "Deep neural network-based noise estimation for robust ASR in dual-microphone smartphones," in *Proc. IberSpeech*, 2016, pp. 117–127.

[14] J. M. Martín-Doñas, A. M. Gomez, I. López-Espejo, and A. M. Peinado, "Dual-channel DNN-based speech enhancement for smartphones," in *Proc. IEEE Workshop on Multimedia Signal Processing (MMSP)*, 2017, pp. 1–6.

[15] K. Tan, X. Zhang, and D. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in *Proc. ICASSP*, 2019, pp. 5751–5755.

[16] J. M. Martín-Doñas, I. López-Espejo, A. M. Gomez, and A. M. Peinado, "An extended Kalman filter for RTF estimation in dual-microphone smartphones," in *Proc. EUSIPCO*, 2018, pp. 2488–2492.

[17] J. M. Martín-Doñas, I. López-Espejo, A. M. Gomez, and A. M. Peinado, "A postfiltering approach for dual-microphone smartphones," in *Proc. IberSpeech*, 2018, pp. 142–146.

[18] J. M. Martín-Doñas, A. M. Peinado, I. López-Espejo, and A. Gomez, "Dual-channel speech enhancement based on extended Kalman filter relative transfer function estimation," *Applied Sciences*, vol. 9, no. 12, p. 2520, 2019.

[19] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, 2016, pp. 196–200.

[20] ——, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech and Language*, vol. 46, pp. 374–385, 2017.

[21] Y. Liu, A. Ganguly, K. Kamath, and T. Kristjansson, "Neural network based time-frequency masking and steering vector estimation for two-channel MVDR beamforming," in *Proc. ICASSP*, 2018, pp. 6717–6721.

[22] S. Chakrabarty and E. Habets, "Time-frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, 2019.

[23] M. Souden, J. Benesty, S. Affes, and J. Chen, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2159–2169, 2011.

[24] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[25] A. Kuklasinski, S. Doclo, S. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 9, pp. 1595–1608, 2016.

[26] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement systems," in *Proc. ICASSP*, 2009, pp. 4409–4412.

[27] J. Heitkaemper, J. Heymann, and R. Haeb-Umbach, "Smoothing along frequency in online neural network supported acoustic beamforming," in *Speech Communication; 13th ITG-Symposium*, 2018.

[28] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. ICASSP*, 2018, pp. 1–5.

[29] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, p. 16, 1988.

[30] L. Lamel, R. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Proc. of the DARPA Speech Recognition Workshop*, 1989, pp. 2161–2170.

[31] D. P. Kingma and J. L. Ba, "ADAM: A method for stochastic optimization," in *Proc. of 3rd International Conference on Learning Representations*, 2015, pp. 1–13.

[32] L. Prechelt, "Early Stopping - But When?" in *Neural Networks: Tricks of the Trade*. Springer Berlin Heidelberg, 2012, pp. 53–67.

[33] "P.862.2: Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codec," ITU-T Std. P.862.2, 2007.

[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[35] J. Roux, S. Wisdom, H. Erdogan, and J. Hershey, "SDR - Half-baked or Well Done?" in *Proc. ICASSP*, 2019, pp. 626–630.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

An analysis of Sound Event Detection under acoustic degradation using multi-resolution systems

Diego de Benito-Gorrón, Daniel Ramos, Doroteo T. Toledano

AUDIAS Research Group
Escuela Politécnica Superior,
Universidad Autónoma de Madrid
(Madrid, Spain)

{diego.benito, daniel.ramos, doroteo.torre}@uam.es

Abstract

The Sound Event Detection task aims to determine the temporal locations of acoustic events in audio clips. Over the recent years, this field is holding a rising relevance due to the introduction of datasets such as Google AudioSet or DESED (Domestic Environment Sound Event Detection) and competitive evaluations like the DCASE Challenge (Detection and Classification of Acoustic Scenes and Events). In this paper, we analyse the performance of Sound Event Detection systems under diverse acoustic conditions such as high-pass or low-pass filtering, clipping or dynamic range compression. For this purpose, the audio has been obtained from the Evaluation subset of the DESED dataset, whereas the systems were trained in the context of the DCASE Challenge 2020 Task 4. Our systems are based upon the challenge baseline, which consists of a Convolutional-Recurrent Neural Network trained using the Mean Teacher method, and they employ a multi-resolution approach which is able to improve the Sound Event Detection performance through the use of several resolutions during the extraction of Mel-spectrogram features. We provide insights on the benefits of this multi-resolution approach in different acoustic settings. Furthermore, we compare the performance of the single-resolution systems in the aforementioned scenarios when using different resolutions.

Index Terms: Sound Event Detection, DCASE Challenge 2020, Multi-resolution, Acoustic degradation.

1. Introduction

Sound events can be defined as the acoustic signals that are directly caused by a particular ocurrence in the near environment, so that a human can identify the event by hearing them. Some clear examples of sound events would be an alarm bell, a dog barking or a person speaking.

Aiming to automatically localize and classify the sound events in audio signals, the task of Sound Event Detection (SED) is an ongoing challenge for machine perception. Training deep learning algorithms in order to develop SED systems requires the use of a large amount of annotated data, which is usually costly to obtain. However, several public datasets have been released over the last years, such as Google AudioSet [1], FSD (FreesoundDataset) [2] or DESED (Domestic Environment Sound Event Detection) [3], which are specifically built to train and evaluate SED systems and consist of audio recordings extracted from web sources, such as YouTube[1], Freesound[2] or Vimeo[3]. These audio recordings can be strongly or weakly annotated, depending on whether the temporal location (onset and offset times) of each event is included or not. In addition to a large scale weakly-labeled audio dataset, Google AudioSet introduced an ontology of more than 500 sound event categories in which sound events can be classified, which has been used as well in the other two mentioned datasets.

On the other hand, the recent development of SED systems and techniques has been notably supported by the DCASE (Detection and Classification of Acoustic Scenes and Events) yearly evaluations, which have helped to define benchmarks not only for Sound Event Detection, but also for other related tasks such as Acoustic Scene Classification [4] or Anomalous Sound Detection [5], among others. Regarding Sound Event Detection, the DCASE 2020 Challenge proposed the task called "Sound event detection and separation in domestic environments", which consisted on locating the temporal boundaries of sound events in ten-seconds audio clips and classifying them.

During the DCASE 2020 Challenge, we developed a multi-resolution approach to Sound Event Detection that was able to outperform the evaluation baseline exploiting the use of several time-frequency resolutions in the process of mel-spectrogram feature extraction, combining up to five different resolution points.

In this paper, we offer an analysis of the performance of single-resolution and multi-resolution SED systems when facing adverse acoustic scenarios that critically affect the spectra of the acoustic signals (high-pass and low-pass filtering) or their dynamic range (clipping and dynamic range compression). For this purpose, we process the audio segments of the Public Evaluation set of DESED in order to achieve the mentioned acoustic conditions, then SED metrics are computed over the obtained sets. Through this study, we aim to determine whether the improvement on performance obtained by the multi-resolution approach is robust to the proposed types of acoustic degradation. These adverse settings represent possible scenarios that could be found when applying the detectors in other data, obtained from web sources or from a real life application.

The rest of the paper is organized as follows: Section 2 explains the Sound Event Detection task of the DCASE Challenge 2020, as well as our multi-resolution approach. Section 3 describes the motivation of this analysis and the different acoustic scenarios that we are considering. In Section 4, the results of the experiments are provided and discussed. Finally, the conclusions of this work are highlighted in Section 5.

[1] http://youtube.com/
[2] http://freesound.org/
[3] http://vimeo.com/

10.21437/IberSPEECH.2021-8

2. Sound Event Detection in DCASE 2020

2.1. DCASE 2020 Challenge: "Sound Event Detection and Separation in Domestic Environments"

In the 2020 edition of the DCASE Challenge, one of the task proposes a Sound Event Detection scenario where systems are trained using the DESED dataset. This dataset includes weakly-labeled data and unlabeled data extracted from Google AudioSet, along with strongly-labeled data which is synthetically generated using the Scaper toolkit [6]. In addition, a subset of AudioSet segments is provided (DESED Validation set) with strong, human-verified annotations, which is used to validate the performance of the systems. An optional pre-processing step based on sound separation is proposed in the task, although our work does not take it into account.

The set of target categories includes ten event categories which are usually found in the acoustic context of a house: *Speech*, *Dog*, *Cat*, *Alarm/bell/ringing*, *Dishes*, *Frying*, *Blender*, *Running water*, *Vacuum cleaner* and *Electric shaver/toothbrush*.

Systems output the predicted onset and offset times of the detected events, along with their category. To define whether a prediction is correct, a collar of 200 ms is considered for the onset times, whereas for the offset times the collar is the maximum between 200 ms and the 20% of the event length, aiming to handle the difficulty to determine the offset times of long events. The system performance is measured by means of the F_1 score metric, which is computed as a combination of the True Positive (TP), False Positive (FP) and False Negative (FN) counts [7].

$$F_1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{1}$$

First, F_1 scores are computed for each event category, then the Macro F_1 is obtained by averaging the class-wise F_1 scores. Macro F_1 is used to measure the global performance of the systems.

The challenge provides a baseline system as a benchmark of SED performance [8]. Such system is based on a Convolutional Recurrent Neural Network trained using the Mean Teacher method [9] for semi-supervised learning. This method allows the network to learn from labeled and unlabeled data. Additionally, an attention module is used to infer the temporal locations of events using weak labels. The system is fed with the mel-spectrogram features of the audio segments.

2.2. Multi-resolution analysis

Each sound event category shows different temporal and spectral characteristics. Therefore, in order to improve SED performance, our main idea is that different time-frequency resolutions would be more suited to detect different types of events. Thus, combining the information of several mel-spectrogram features extracted at different resolution points should lead to a better overall performance [10].

Aiming to test this hypothesis, we defined five different time-frequency resolutions, taking as a starting point the resolution used by the baseline system, which we called BS. Each resolution point is defined by a set of values for the parameters of feature extraction. It should be noted that, due to the feature extraction process, there is a compromise between temporal resolution and frequency resolution. Hence, we propose a resolution point with twice better time resolution than the baseline, which we call T_{++}, and a resolution point with twice better frequency resolution than the baseline, which we call F_{++}. In the intermediate points between each of these points and BS, we define T_{+} and F_{+}, respectively.

In order to obtain multi-resolution systems, first we trained single-resolution systems, which were based on the DCASE Challenge baseline and modified to operate on each of the different resolution points. Then, we performed a model fusion averaging the posterior probabilities given by systems trained with different resolutions. Using this method, we obtained a three-resolution system which combines the BS resolution with T_{++} and F_{++}, denoted as *3res* in this paper, and a five-resolution system combining all the mentioned resolutions, denoted as *5res*.

Through the use of the *3res* and *5res* systems, we were able to outperform the single-resolution baseline system in the DCASE 2020 Challenge task 4. The improvement of performance in terms of macro F_1 score was observed over the DESED Validation set and the YouTube subset of the DESED 2019 Evaluation set, which is called DESED Public evaluation set. The *5res* system was submitted to the evaluation and outperformed the baseline system over the DESED 2020 Evaluation set.

3. Experiments

Both the DESED Validation set and the Public Evaluation set consist of YouTube audio segments drawn from Google AudioSet. Due to the crowdsourced nature of a web resource like YouTube, the audio clips can have very diverse origins and qualities, ranging from mobile recordings to professional studio productions. Therefore, the evaluation of Sound Event Detection on YouTube data requires the systems to be able to handle a variety of acoustic conditions that sometimes may be adverse for the task.

In order to test the performance of Sound Event Detection in a wider range of acoustic settings, we have applied several types of degradations to the DESED Public evaluation set, which contains 692 audio clips. We have computed the F_1 scores of single-resolution and multi-resolution systems over the original set and its degraded copies, aiming to analyze to what extent does multi-resolution help to improve performance when the test data is degraded.

The acoustic conditions that we have considered for our experiments are inspired by some of the scenarios described in the DESED Synthetic evaluation set, which has already been used to analyze the performance of state-of-the-art SED systems [11].

3.1. Acoustic degradation scenarios

We consider three types of degradations for the audio clips: frequency filtering, dynamic range compression and clipping. We apply each perturbation to the whole dataset, obtaining a total of eight copies of the DESED Public evaluation set:

- **Frequency filtering.** We apply high-pass filtering and low-pass filtering separately. In both cases, the cutoff frequencies are 500 Hz, 1000 Hz and 2000 Hz, leading to a total of six copies of the DESED Public evaluation set.
- **Dynamic range compression.** We apply dynamic range compression with a threshold of -50 dB and a ratio value of 5.
- **Clipping.** To obtain clipping distortion, we multiply the

audio signals, which are bounded to $[-1, 1]$, by a scale factor of 5, limiting the output values again to $[-1, 1]$.

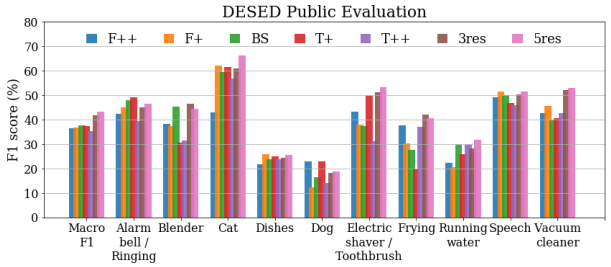


Figure 1: *F_1 scores of the single-resolution systems ($F_{++}, F_{+}, BS, T_{+}, T_{++}$) and the multi-resolution systems (3res, 5res) over the DESED Public Evaluation set. Best viewed in color.*

4. Results

All the results are provided in terms of event-based F_1 score, considering the same collar settings as in the DCASE 2020 Challenge task 4.

4.1. Results over DESED Public evaluation set

The results of the seven systems over the original DESED Public evaluation set are presented in Figure 1. The figure represents the F_1 scores of each system in groups of bars, one group for each event category and an additional one for the macro average, which represents the global performance.

It can be observed that, in terms of macro F_1, the *3res* and *5res* systems both outperform every single-resolution system. However, this improvement is not applicable to every target class. Whereas most event categories obtain their best performance when using a multi-resolution system, other classes reach their maximum F_1 score with a single-resolution system: This is the case of *Alarm bell/ringing* (T_{+}), *Dishes* (F_{+}), *Dog* (T_{+}) and *Speech* ($F+$).

4.2. Results under acoustic degradation

4.2.1. High-pass filtering

The results obtained when applying high-pass filtering to the DESED Public evaluation set are shown in Figure 2. Three separate graphs are presented, one for each cutoff frequency (f_c). As expected, the general performance decreases for every class and every system when the cutoff frequency of the high-pass filter increases. In terms of macro F_1 score, the multi-resolution systems *3res* and *5res* achieve the best results for $f_c = 500$ Hz, similarly to the clean set. However, for $f_c = 1000$ Hz and $f_c = 2000$ Hz the highest macro F_1 scores are obtained with some of the single-resolution systems, BS and T_{+}, respectively.

4.2.2. Low-pass filtering

Figure 3 shows the results for the DESED Public evaluation set after applying low-pass filtering with $f_c = 2000$ Hz, $f_c = 1000$ Hz and $f_c = 500$ Hz. It can be seen that the performances decrease when lowering the cutoff frequency of the filter, which is the expected behavior. When using a cutoff frequency $f_c =$

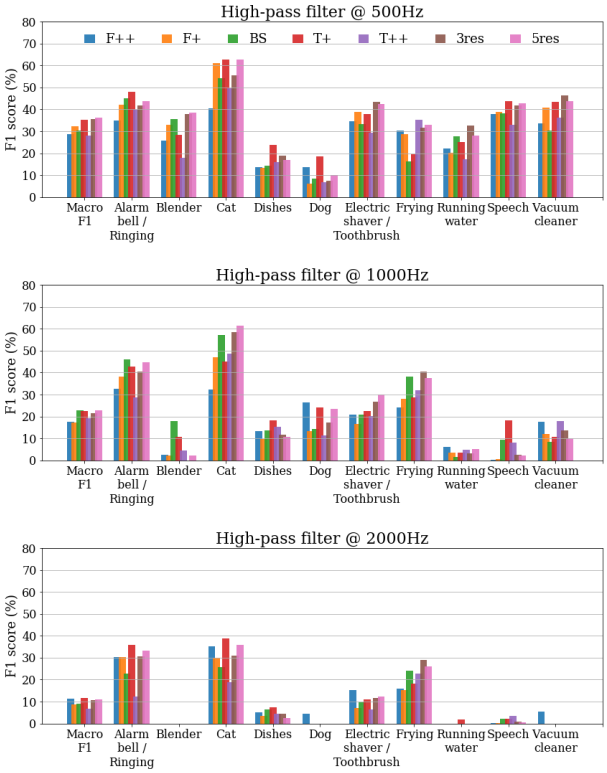


Figure 2: *F_1 scores of the single-resolution systems ($F_{++}, F_{+}, BS, T_{+}, T_{++}$) and the multi-resolution systems (3res, 5res) over the DESED Public Evaluation set applying a high-pass filter with cutoff frequencies of 500 Hz (top), 1000 Hz (center) and 2000 Hz (bottom). Best viewed in color.*

2000 Hz, the best overall performance is obtained by the multi-resolution system *5res*, whereas for $f_c = 1000$ Hz *3res* and T_{++} both achieve the highest macro F_1. When the cutoff frequency is set to $f_c = 500$ Hz, the best macro F_1 scores are obtained with the single-resolution systems T_{+} and T_{++}.

4.2.3. Dynamic range compression

The results obtained after applying dynamic range compression to the DESED Public evaluation set are presented in Figure 4. In this scenario, the best overall performance (macro F_1) is obtained by the multi-resolution systems, *3res* and *5res*. However, for some particular classes the best performance is obtained with a single-resolution system, as observed in the clean set results. Such is the case of *Alarm bell/ringing* (F_{++}), *Cat* (F_{+}), *Dishes* (T_{++}) and *Running water* (T_{++}).

4.2.4. Clipping

Figure 5 presents the results obtained when applying clipping saturation to the Public evaluation set. The best macro F_1 performances are achieved by the multi-resolution systems, whereas in some event categories multi-resolution is not able to outperform every single-resolution system. This situation is observed for *Dishes* (T_{+}), *Dog* (F_{++}) and *Shaver* (F_{+}).

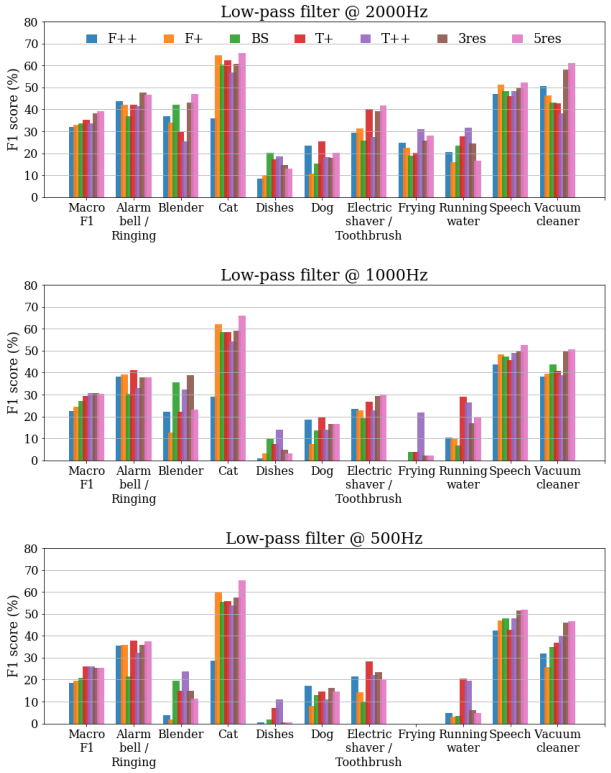


Figure 3: *F_1 scores of the single-resolution systems $(F_{++}, F_{+}, BS, T_{+}, T_{++})$ and the multi-resolution systems (3res, 5res) over the DESED Public Evaluation set applying a low-pass filter with cutoff frequencies of 2000 Hz (top), 1000 Hz (center) and 500 Hz (bottom). Best viewed in color.*

4.3. Discussion

It has been shown that the proposed adverse settings have a negative impact on the performance of both single-resolution and multi-resolution Sound Event Detection systems. Overall, the most critical scenario is high-pass filtering, especially with f_c values of 1000 Hz and above. This suggests that the information of low frequencies is essential for this task, especially when considering categories like *Blender*, *Speech*, *Running water* or *Vacuum cleaner*. On the other hand, low-pass filtering is the most adverse condition for the class *Frying*, implying that high frequencies are particularly relevant for this event.

The results also show that the improvement on performance obtained when combining several single-resolution systems into a multi-resolution system does not always hold when facing very adverse conditions. Likely, this effect is due to the way in which our multi-resolution systems are obtained. An average fusion of the scores of different models can result in more accurate scores when the individual scores are precise enough. On the other hand, in scenarios where the individual systems perform worse, the average fusion is not able to obtain better results.

Nevertheless, it can be observed that, under these adverse settings, multi-resolution systems have an overall result approximately as good as the best performing resolution in each case, which means that our multi-resolution approach provides an improved robustness against these very adverse distortion scenarios.

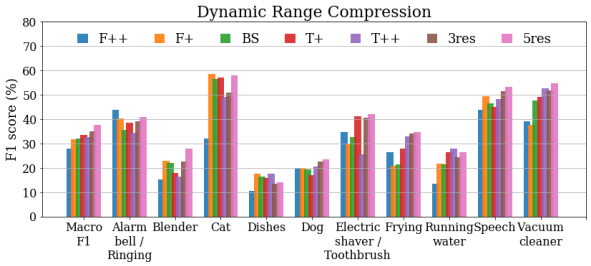


Figure 4: *F_1 scores of the single-resolution systems $(F_{++}, F_{+}, BS, T_{+}, T_{++})$ and the multi-resolution systems (3res, 5res) over the DESED Public Evaluation set applying dynamic range compression. Best viewed in color.*

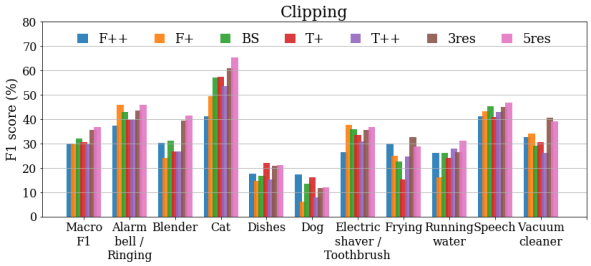


Figure 5: *F_1 scores of the single-resolution systems $(F_{++}, F_{+}, BS, T_{+}, T_{++})$ and the multi-resolution systems (3res, 5res) over the DESED Public Evaluation set applying clipping saturation. Best viewed in color.*

5. Conclusions

In this paper, we have studied the performance of several Sound Event Detection systems over a public dataset when diverse acoustic perturbations are applied. Five of these systems are convolutional neural networks with a common structure, but employing mel-spectrogram features extracted using different time-frequency resolutions. Two more systems are considered, which combine the previous systems into multi-resolution models by means of an average fusion, increasing the performance over the evaluation subsets of the DESED dataset.

According to the results, the proposed acoustic scenarios have, as expected, a clearly negative impact on the performance of our systems. Although it is shown that our multi-resolution approach is robust to slight degradations, the average fusion is unable to improve performance when facing very adverse conditions. Additionally, an extra robustness against these adverse distortion scenarios is observed when using multiple resolutions. We are currently working on alternative implementations of the multi-resolution approach in which fusion is performed earlier, aiming to improve both performance and robustness.

Furthermore, the data generated and the results obtained through this study will serve as a benchmark to evaluate the performance of future Sound Event Detection approaches and their robustness to diverse acoustic settings.

6. Acknowledgements

Work developed under project DSForSec (RTI2018-098091-B-I00), funded by the Ministry of Science, Innovation and Universities of Spain and FEDER.

7. References

[1] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[2] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*, Suzhou, China, 2017, pp. 486–493.

[3] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, 2020. [Online]. Available: https://hal.inria.fr/hal-02355573

[4] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in DCASE 2020 Challenge: generalization across devices and low complexity solutions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, submitted. [Online]. Available: https://arxiv.org/abs/2005.14623

[5] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and Discussion on DCASE2020 Challenge Task2: Unsupervised Anomalous Sound Detection for Machine Condition Monitoring," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 81–85.

[6] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 344–348.

[7] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, May 2016. [Online]. Available: http://dx.doi.org/10.3390/app6060162

[8] N. Turpault and R. Serizel, "Training sound event detection on a heterogeneous dataset," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 200–204.

[9] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[10] D. de Benito-Gorron, D. Ramos, and D. T. Toledano, "A multi-resolution approach to sound event detection in DCASE 2020 task4," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 36–40.

[11] N. Turpault, R. Serizel, S. Wisdom, H. Erdogan, J. Hershey, E. Fonseca, P. Seetharaman, and J. Salamon, "Sound Event Detection and Separation: a Benchmark on DESED Synthetic Soundscapes," 2020. [Online]. Available: arXiv:2011.00801 [cs.SD]

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Speech Enhancement for Wake-Up-Word detection in Voice Assistants

David Bonet[1,5], *Guillermo Cámbara*[2,5], *Fernando López*[4],
Pablo Gómez[4], *Carlos Segura*[5], *Mireia Farrús*[3], *Jordi Luque*[5]

[1]Universitat Politècnica de Catalunya, [2]Universitat Pompeu Fabra,
[3]Universitat de Barcelona, [4]Telefónica I+D, Digital Home
[5]Telefónica I+D, Research, Spain

jordi.luque@telefonica.com

Abstract

Keyword spotting and in particular Wake-Up-Word (WUW) detection is a very important task for voice assistants. A very common issue of voice assistants is that they get easily activated by background noise like music, TV or background speech that accidentally triggers the device. In this paper, we propose a Speech Enhancement (SE) model adapted to the task of WUW detection that aims at increasing the recognition rate and reducing the false alarms in the presence of these types of noises. The SE model is a fully-convolutional denoising auto-encoder at waveform level and is trained using a log-Mel Spectrogram and waveform reconstruction losses together with the BCE loss of a simple WUW classification network. A new database has been purposely prepared for the task of recognizing the WUW in challenging conditions containing negative samples that are very phonetically similar to the keyword. The database is extended with public databases and an exhaustive data augmentation to simulate different noises and environments. The results obtained by concatenating the SE with a simple and state-of-the-art WUW detectors show that the SE does not have a negative impact on the recognition rate in quiet environments while increasing the performance in the presence of noise, especially when the SE and WUW detector are trained jointly end-to-end.
Index Terms: keyword spotting, speech enhancement, wake-up-word, deep learning, convolutional neural network

1. Introduction

Voice interaction with devices is becoming ubiquitous. Most of them use a mechanism to avoid the excessive usage of resources, a trigger word detector. This ensures the efficient use of resources, using a Speech-To-Text tool only when needed and with the consequent start of a conversation. It is key to only start this conversation when the user is addressing the device, otherwise the user experience is notoriously degraded. Thus, the wake-up-word detection system must be robust enough to avoid wake-ups with TV, music, speech and sounds that do not contain the key phrase.

A common approach to reduce the impact of this type of noise in the system is the adoption of speech enhancement algorithms. Speech enhancement consists of the task of improving the perceptual intelligibility and quality of speech by removing background noise [1]. Its main application is in the field of mobile and internet communications [2] and related to hearing aids [3], but SE has also been applied successfully to automatic speech recognition systems [4, 5, 6].

Traditional SE methods involved a characterization step of the noise spectrum which is then used to try reduce the noise from the regenerated speech signal. Examples of these approaches are spectral subtraction [3], Wiener filtering [7] and subspace algorithms [8]. One of the main drawbacks of the classical approaches is that they are not very robust against non-stationary noises or other type of noises that can mask speech, like background speech. In the last years, Deep Learning approaches have been widely applied to SE at the waveform level [9, 10] and spectral level [6, 11]. In the first case, a common architecture falls within the encoder-decoder paradigm. In [12], authors proposed a fully convolutional generative adversarial network architecture structured as an auto-encoder with U-Net like skip-connections. Other recent work [13] proposes a similar architecture at the waveform level that includes a LSTM between the encoder and the decoder and it is trained directly with a regression loss combined with a spectrogram domain loss.

Inspired by these recent models, we propose a similar SE auto-encoder architecture in the time domain that is optimized not only by minimizing waveform and Mel-spectrogram regression losses, but also includes a task-dependent classification loss provided by a simple WUW classifier acting as a Quality-Net [14, 15]. This last term serves as a task-dependent objective quality measure that trains the model to enhance important speech features that might be degraded otherwise. The WUW detection is performed by concatenating the SE model with the classifier.

2. Speech Enhancement

Speech enhancement is interesting for triggering phrase detection since it tries to remove noise that could trigger the device, and at the same time improves speech quality and intelligibility for a better detection. In this case, we try to tackle the most common noisy environments where voice assistants are used: TV, music, background conversations, office noise and living room noise. Some of these types of background noise, such as TV and background conversations, are the most likely to trigger the voice assistant and are also the most challenging to remove.

2.1. Model

Our model has a fully-convolutional denoising auto-encoder architecture with skip connections (Fig. 1), working end-to-end at waveform level. Similar designs have proven to be very effective in SE tasks [12, 13, 16]. In training, we input a noisy audio $\boldsymbol{x} \in \mathbb{R}^T$, comprised of clean speech signal $\boldsymbol{y} \in \mathbb{R}^T$ and background noise $\boldsymbol{n} \in \mathbb{R}^T$ so that $\boldsymbol{x} = \lambda \boldsymbol{y} + (1 - \lambda)\boldsymbol{n}$.

The encoder compresses the input signal and expands the number of channels. It is composed of six convolutional blocks (ConvBlock1D), each consisting of a convolutional layer, followed by an instance normalization and a rectified linear unit (ReLU). Kernel size $K = 4$ and stride $S = 2$ are used, except in the first layer where $K = 7$ and $S = 1$. The compressed sig-

10.21437/IberSPEECH.2021-9

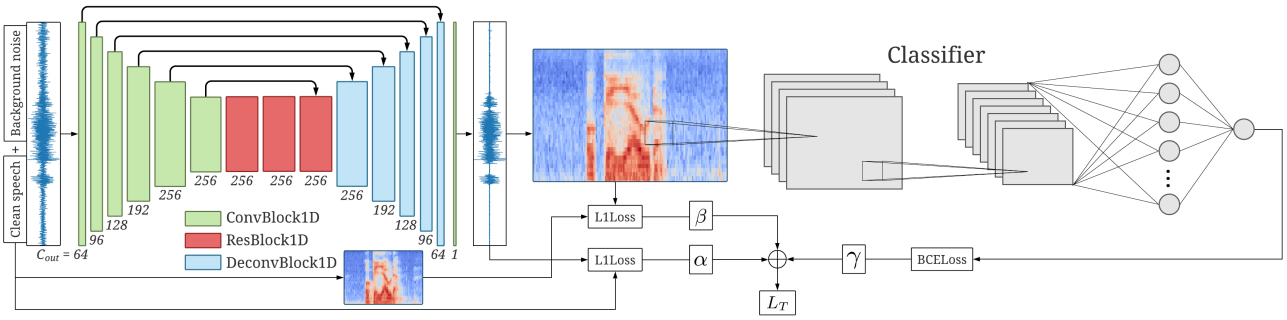


Figure 1: *End-to-end SE model at waveform level concatenated with a classifier. The Log-Mel Spectrogram and waveform reconstruction losses of the SE model can be used together with the task-dependent loss (BCE Loss) of the classifier acting as a Quality-Net [15] to train the model. The latter term aims at enhancing relevant speech features for the WUW detection task.*

nal goes through an intermediate stage where the shape is preserved, consisting of three residual blocks (ResBlock1D), each formed by two ConvBlock1D with $K = 3$ and $S = 1$ where a skip connection is added from the input of the residual block to its output. The last stage of the SE model is the decoder, where the original shape of the raw audio is recovered at the output, and the enhanced signal can serve as input to a WUW classifier. Its architecture follows the inverse structure of the encoder, where deconvolutional blocks (DeconvBlock1D) replace the convolutional layers of the ConvBlock1D with transposed convolutional layers. Skip connections from the encoder blocks to the decoder blocks are also used to ensure low-level detail when reconstructing the waveform.

We use a regression loss function (L1 loss) at raw waveform level together with another L1 loss over the log-Mel Spectrogram as proposed in [17] to reconstruct a "cleaned" signal $\hat{\boldsymbol{y}}$ at the output. Finally, we include the classification loss (BCE Loss) when training the SE model jointly with the classifier or concatenating a pretrained classifier at its output. Thus, we also try to optimize the SE model to the specific task of WUW classification. Our final loss function is defined as a linear combination of the three losses:

$$L_T = \alpha L_{raw}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \beta L_{spec}(S(\boldsymbol{y}), S(\hat{\boldsymbol{y}})) + \gamma L_{BCE} \quad (1)$$

where α, β and γ are hyperparamters weighting each loss term and $S(\cdot)$ denotes the log-Mel Spectrogram of the signal, which is computed using 512 FFT bins, a window of 20 ms with 10 ms of shift and 40 filters in the Mel Scale.

3. Methodology

3.1. Databases

The database used for conducting the experiments here presented consists of WUW samples labeled as positive, and other non-WUW samples labeled as negative. Since the chosen keyword is "OK Aura", which triggers Telefónica's home assistant, Aura, positive samples are drawn from company's in-house databases. Some of the negative samples have been also recorded in such databases, but we also add speech and other types of acoustic events from external data sets, so the models gain robustness with further data augmentation. Information about all data used is detailed in this section.

3.1.1. OK Aura Database

In a first round, around 4300 WUW samples from 360 speakers have been collected, resulting in 2.8 hours of audio. Furthermore, office ambient noise has been recorded as well, with the aim of having samples for noise data augmentation. The second data collection round has been done in order to study and improve some sensitive cases where WUW modules typically underperform. For instance, such dataset contains rich metadata about positive and negative utterances, like room distance, speech accent, emotion, age or gender. Furthermore, the negative utterances contain phonetically similar words to "OK Aura", since these are the most ambiguous to recognize for a classifier. Detailed information about data acquisition is explained in the following subsection.

3.1.2. Data acquisition

A web-based Jotform form[1] has been designed for data collection. Readers are invited to contribute to the dataset while the form is still open. Until the date of this work, 1096 samples from 80 speakers have been recorded, which consists of 1.2 hours of audio[2]. Volunteers are asked to pronounce various scripted utterances at a close distance and also at two meters from the device mic. The similarity levels are the following:

1. Exact WUW, in an isolated manner: *OK Aura.*
2. Exact WUW, in a context: *Perfecto, voy a mirar qué dan hoy. OK Aura.*
3. Contains "Aura": *Hay un aura de paz y tranquilidad.*
4. Contains "OK": *OK, a ver qué ponen en la tele.*
5. Contains similar word units to "Aura": *Hola Laura.*
6. Contains similar word units to "OK": *Prefiero el hockey al baloncesto.*
7. Contains similar word units to "OK Aura": *Porque Laura, ¿qué te pareció la película?*

3.1.3. External data

General negative examples have been randomly chosen from the publicly available Spanish Common Voice (CV) corpus [18] that currently holds over 300 hours of validated audio. However, we keep a 10:1 ratio between negative and positive samples, since such ratio proves to yield good results in [19], thus avoiding bigger ratios that lead to increasing computational times. Final database collects a CV partition consisting of 55h for training, 7h for development and 7h for testing.

Background noises were selected from various public datasets according to different use case scenarios. Living room

[1] https://form.jotform.com/201694606537056

[2] The AURA-WUW dataset, including audio and alignments, is available upon request from the authors and agreement of EULA for research purposes.

background noise (HOME-LIVINGB) from the QUT-NOISE Database [20], TV audios from the IberSpeech-RTVE Challenge [21], and music[3] and conversations[4] from free libraries.

3.1.4. Data processing

All the audio samples are monoaural signals stored in Waveform Audio File Format (WAV) with a sampling rate of 16kHz. The speech data that has been collected was processed with a Speech Activity Detection (SAD) module producing timestamps where speech occurs. For this purpose the tool from pyannote.audio [22] has been used, which has been trained with the AMI corpus [23]. This helped us to only use the valid speech segments of the audios we collected.

As features to train the classifiers we mainly used two, as we wanted to maintain the architectures with the originally proposed features [24, 25]: Mel-Frequency Cepstral Coefficients (MFCCs) and log-Mel Spectrogram. The MFCCs were constructed first filtering the audio with a band pass filter (20Hz to 8kHz) and then, extracting the first thirteen coefficients with 100 ms of windows size and frame shifting of 50 ms. The procedure to extract the log-Mel Spectrogram ($S(\cdot)$) is detailed in §2.1.

Train, development and test partitions are split ensuring that neither speaker nor background noise is repeated between partitions, trying to maintain a 80-10-10 proportion, respectively. Total data, containing internal and external datasets, consists of 50.737 non-WUW samples and 4.651 WUW samples.

3.2. Data augmentation

Several Room Impulse Responses (RIR) were created based on the Image Source Method (ISM) [26], for a room of dimensions (L_x, L_y, L_z) where $2 \leq L_x \leq 4.5, 2 \leq L_y \leq 5.5, 2.5 \leq L_z \leq 4$ meters, with microphone and source randomly located at any (x, y) point within a height of $0.5 \leq z \leq 2$ meters. Every TV and music original recordings were convolved with different RIRs to simulate the signal picked up by the microphone of the device in the room.

The main data augmentation technique used in this work is background noise addition. Different noise scenarios, like TV, music, conversations, office and living room, have been combined with clean samples within a wide range of SNRs. It aims at improving the performance of the models against noisy environments. In each epoch, we create different noisy samples by randomly selecting a sample of background noise for each speech event and combining them with a randomly chosen SNR in a specified range. We tried data augmentation techniques like time stretching and pitch shifting, but we discarded them since no significant changes were achieved in the noisy regions.

3.3. Wake-Up Word Detection Models

With the aim of assessing the quality of the trained SE models, we use several trigger word detection classifier models, reporting the impact of the SE module at WUW classification performance. The WUW classifiers used here are a `LeNet`, a well-known standard classifier, easy to optimize [27]; `Res15`, `Res15-narrow` and `Res8` based on a reimplementation by Tang and Lin [28] of Sainath and Parada's Convolutional Neural Networks (CNNs) for keyword spotting [29], using residual learning techniques with dilated convolutions [30]; a `SGRU` and

[3]https://freemusicarchive.org/

[4]http://www.podcastsinspanish.org/

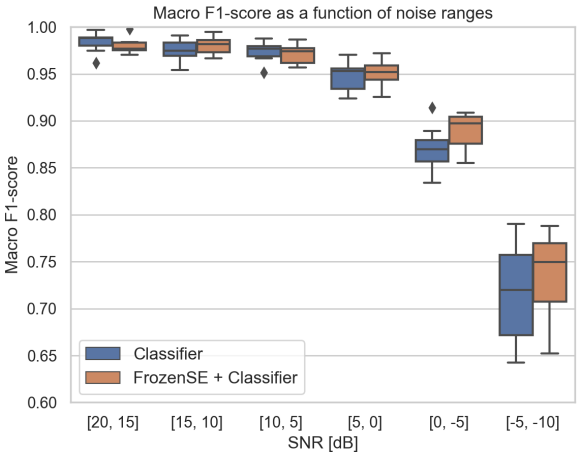


Figure 2: *Macro F1-score box plot for different SNR ranges. Classifiers trained with low noise ($[5, 30]$ dB SNR).*

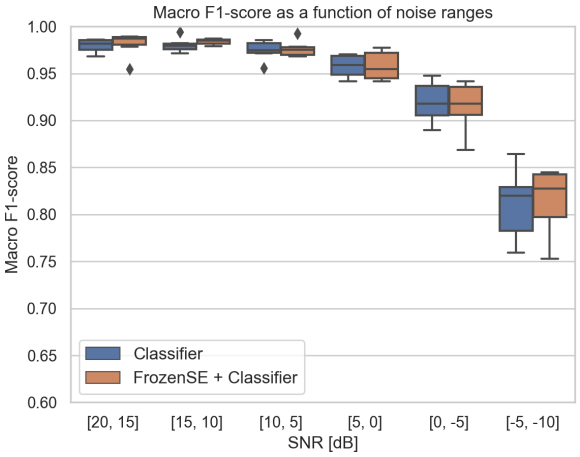


Figure 3: *Macro F1-score box plot for different SNR ranges. Classifiers trained with a very wide range of noise ($[-10, 50]$ dB SNR).*

`SGRU2`, two Recurrent Neural Network (RNNs) models, based on the open source tool named Mycroft Precise [24], which is a lightweight wake-up-word detection tool implemented in TensorFlow. These are two bigger variations that we have implemented in PyTorch. We also use a `CNN-FAT2019`, a CNN architecture adapted from a kernel [25] in Kaggle's FAT 2019 competition [31], which has shown good performance in tasks like audio tagging or detection of gender, identity and speech events from pulse signal [32].

3.4. Training

Speech signals and background noises are combined randomly following the procedure explained in 3.2 with a given SNR range. The SE model is trained to cover a wide SNR range of $[-10, 50]$ dBs, whereas WUW models are trained to cover two scenarios: a classifier trained with the same SNR range as the SE model, and a classifier less aware of noise with a narrower SNR range of $[5, 30]$ dBs. This way, it is possible to study the impact of the SE model regarding if the classifier has been trained with more or less noise.

Data imbalance is addressed balancing the classes in each batch using a weighted sampler. We use a fixed window length of 1.5 s based on the annotated timestamps for our collected database, and random cuts for the rest of the CV samples.

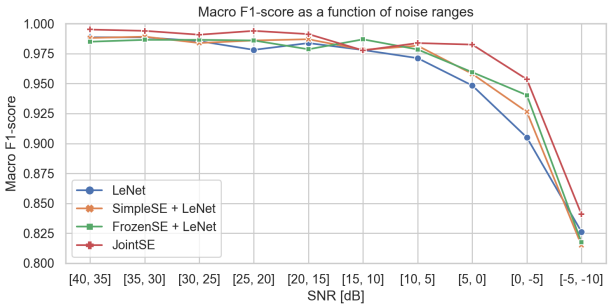


Figure 4: *Comparison of different training methods for the SE models and LeNet classifier, in terms of the macro F1-Score for different SNR ranges. All models trained in the range of* $[-10, 50] dB$ *SNR.*

All the models are trained with early stopping based on the validation loss with 10 epochs of patience. We use the Adam optimizer with a learning rate of 0.001 and a batch size of 50. Loss (1) allows to train the models in multiple ways and we define different SE models and classifiers based on the loss function used:

a) Classifier: we remove the auto-encoder from the architecture (Fig. 1) and train any of the classifiers using the noisy audio as input: $\alpha = \beta = 0$ and $\gamma = 1$

b) SE model (SimpleSE): we remove the classifier from the architecture and optimize the auto-encoder based on the reconstruction losses only: $\alpha = \beta = 1$ and $\gamma = 0$

c) SE model + frozen classifier (FrozenSE): operations of the classifier are dropped from the backward graph for gradient calculation, optimizing only the SE model for a given pretrained classifier (LeNet). $\alpha = \beta = \gamma = 1$

d) SE model + classifier (JointSE): auto-encoder and LeNet are trained jointly using the three losses: $\alpha = \beta = \gamma = 1$

3.5. Tests

All the models take as input windows of 1.5 s of audio, to ensure that common WUW utterances are fully within it, since the average "OK Aura" is about 0.8 s long. Therefore, we perform an atomic test evaluating if a single window contains the WUW or not. Both negative and positive samples are assigned a background noise sample with which they are combined with a random SNR between certain ranges, as described in §3.4.

Given the output scores of the models, the decision threshold is chosen as the one yielding the biggest difference between true and false positive rates, based on Youden's J statistic [33]. Once the threshold is decided, macro F1-score is computed in order to balance WUW/non-WUW proportions in the results. We average such scores across all WUW classifiers described in §3.3, for every SNR range.

4. Results

Figure 2 illustrates the improvement of the WUW detection in noisy scenarios by concatenating our FrozenSE model with all WUW classifiers described in §3.3 trained with low noise ($[5, 30]$ dB SNR), which we could find in simple voice assistant systems. Applying SE in quiet scenarios maintains fairly good results, and improves them in lower SNR ranges.

If we train the classifiers with more data augmentation ($[-10, 50]$ dB SNR), results using the FrozenSE do not decrease but the improvement in ranges of severe noise is not as large as in Figure 2, see Figure 3.

Table 1: *Macro F1-score enhancing the noisy audios with SOTA SE models and using a LeNet as a classifier.*

SNR $[dB]$		**No SE**	**SEGAN**	**Denoiser**	**JointSE**
$[20, 10]$	Clean	0.980	0.964	0.980	**0.990**
$[10, 0]$	Noisy	0.969	0.940	0.955	**0.972**
$[0, -10]$	Very noisy	0.869	0.798	0.851	**0.902**

Table 2: *Macro F1-score percentage difference between JointSE and LeNet without SE module, for different background noise types. Positive values mean that the JointSE score is bigger than the single LeNet's.*

SNR $[dB]$		**Music**	**TV**	**Office**	**Living Room**	**Conversations**
$[20, 10]$	Clean	1.0	-0.9	1.4	0.4	**2.3**
$[10, 0]$	Noisy	0.0	-1.2	0.8	0.4	**1.9**
$[0, -10]$	Very noisy	0.5	3.9	**11.2**	3.1	3.8

In §3.4 we have defined the parameters of the loss function (1) to train a classifier (case a)), and different approaches to train the SE model, either standalone (b), c)) or in conjunction with the classifier (d)). In Figure 4 we can see how JointSE performs better than all the other cases in almost every SNR range. From 40 dB to 10 dB of SNR, the results are very similar for the 4 models. In contrast, in the noisiest ranges we can see how the classifier without SE model is the worst performer, followed by the SimpleSE case where only the waveform and spectral reconstruction losses are used. We found that the FrozenSE case, which includes the classification loss in the training stage, improves the results for the wake-up-word detection task. However, the best results are obtained with the JointSE case where the SE model + LeNet are trained jointly using all three losses.

We compared the WUW detection results of our JointSE with other SOTA SE models (SEGAN [12] and Denoiser [13]), followed by a classifier (data augmented LeNet) in different noise scenarios. In Table 1, it can be observed how when training the models together with the task loss, the results in our setup are better than with other more powerful but more general SE models, since there is no mismatch between the SE and classifier in the end-to-end and it is also more adapted to common home noises. JointSE improves the detection over the no SE model case, especially in scenarios with background conversations, loud office noise or loud TV, see Table 2.

5. Conclusions

In this paper we proposed a SE model adapted to the task of WUW in voice assistants for the home environment. The SE model is a fully-convolutional denoising auto-encoder at waveform level and it is trained using a log-Mel Spectrogram and waveform regression losses together with a task-dependent WUW classification loss. Results show that for clean and slightly noisy conditions, SE in general does not bring a substantial improvement over a classifier trained with proper data augmentation. In very noisy conditions, SE does improve the results, especially when the SE model and WUW detector are trained jointly end-to-end, performing better than a general-purpose SE model.

6. Acknowledgments

This work has been partly funded by the INGENIOUS project within the European Union's Horizon 2020 Research and Innovation Programme under GA No 833435 and by the Agencia Estatal de Investigación (AEI), Ministerio de Ciencia, Innovación y Universidades and the Fondo Social Europeo (FSE) under grant RYC-2015-17239 (AEI/FSE, UE).

7. References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[2] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," *arXiv preprint arXiv:1909.08050*, 2019.

[3] L.-P. Yang and Q.-J. Fu, "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise," *The journal of the Acoustical Society of America*, vol. 117, no. 3, pp. 1001–1004, 2005.

[4] C. Zorilă, C. Boeddeker, R. Doddipatla, and R. Haeb-Umbach, "An investigation into the effectiveness of enhancement in ASR training and test for chime-5 dinner party transcription," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 47–53.

[5] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[6] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[7] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction," in *1997 IEEE international conference on acoustics, speech, and signal processing*, vol. 2. IEEE, 1997, pp. 1167–1170.

[8] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on speech and audio processing*, vol. 3, no. 4, pp. 251–266, 1995.

[9] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[10] H. Phan, I. V. McLoughlin, L. Pham, O. Y. Chén, P. Koch, M. De Vos, and A. Mertins, "Improving GANs for speech enhancement," *arXiv preprint arXiv:2001.05532*, 2020.

[11] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*, 2016.

[12] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[13] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[14] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," *arXiv preprint arXiv:1808.05344*, 2018.

[15] S.-W. Fu, C.-F. Liao, and Y. Tsao, "Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality," *IEEE Signal Processing Letters*, vol. 27, pp. 26–30, 2019.

[16] J. Llombart, D. Ribas, A. Miguel, L. Vicente, A. Ortega, and E. Lleida, "Progressive loss functions for speech enhancement with deep neural networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–16, 2021.

[17] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[18] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[19] J. Hou, Y. Shi, M. Ostendorf, M.-Y. Hwang, and L. Xie, "Mining effective negative training samples for keyword spotting," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7444–7448.

[20] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," *Proceedings of Interspeech 2010*, 2010.

[21] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. de Prada, "Albayzin 2018 evaluation: the IberSpeech-RTVE challenge on speech technologies for Spanish broadcast media," *Applied Sciences*, vol. 9, no. 24, p. 5412, 2019.

[22] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

[23] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[24] M. D. Scholefield, "Mycroft Precise," https://github.com/MycroftAI/mycroft-precise, 2019.

[25] M. H. "mhiro2", "Freesound Audio Tagging 2019: Simple 2D-CNN Classifier with PyTorch," https://www.kaggle.com/mhiro2/simple-2d-cnn-classifier-with-pytorch/, 2019.

[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[27] Y. LeCun *et al.*, "Lenet-5, convolutional neural networks," *URL: http://yann. lecun. com/exdb/lenet*, vol. 20, no. 5, p. 14, 2015.

[28] R. Tang and J. Lin, "Honk: A PyTorch reimplementation of convolutional neural networks for keyword spotting," *CoRR*, vol. abs/1710.06554, 2017. [Online]. Available: http://arxiv.org/abs/1710.06554

[29] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[30] R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," *CoRR*, vol. abs/1710.10361, 2017. [Online]. Available: http://arxiv.org/abs/1710.10361

[31] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, and X. Serra, "Audio tagging with noisy labels and minimal supervision," *arXiv preprint arXiv:1906.02975*, 2019.

[32] G. Cámbara, J. Luque, and M. Farrús, "Detection of speech events and speaker characteristics through photo-plethysmographic signal neural processing," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7564–7568.

[33] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

An approach to intent detection and classification based on attentive recurrent neural networks

Fernando Fernández-Martínez[1], *David Griol*[2], *Zoraida Callejas*[2], *Cristina Luna-Jiménez*[1]

[1]Speech Technology Group, Center for Information Processing and Telecommunications, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid, Spain
[2]Department of Languages and Computer Systems, University of Granada, Spain

fernando.fernandezm@upm.es, dgriol@ugr.es, zoraida@ugr.es, cristina.lunaj@upm.es

Abstract

Intent Detection is a key component of any task-oriented conversational system. To understand the user's current goal and provide the most adequate response, the system must leverage its intent detector to classify the user's utterance into one of several predefined classes (intents). This objective can also simplify the set of processes that a conversational system must complete by performing the natural language understanding and dialog management tasks into a single process conducted by the intent detector. This is particularly useful for systems oriented to FAQ services. In this paper we present a novel approach for intent detection and classification based on word-embeddings and recurrent neural networks. We have validated our approach with a selection of the corpus acquired with the Hispabot-Covid19 system obtaining satisfactory results.

Index Terms: topic classification, intent detection, conversational systems, recurrent networks, attentive rnn, attentive lstm

1. Introduction

Traditional spoken language understanding in conversational systems is divided into two main subtasks, intent detection and semantic slot filling, which are extended with domain recognition in multi-domain dialogue systems [1, 2].

The main objective of intent detection (also known as intent recognition or intent classification) is to classify user utterances into previously defined intent categories according to the domains and intents involved in user utterances [3, 4]. Users' intents can be defined as the will of the user (i.e., what they want to do). They are also denoted as dialogue acts, which can be defined as information actions that users share in the dialogue and are constantly updated (e.g., ask for train schedules, book a hotel, etc.).

This is a complex process that involves several key challenges: the lack of data sources to apply statistical methodologies (there are very few corpora with intents annotations and they are very difficult to obtain) [5], the irregularity of users' expressions (the use of colloquial expression, short sentences and broad content make very difficult to identify user's intents) [6], implicit intent detection (implicit intents are those in which users do not have clear intent requirements and it is necessary to infer the user's real intent by analyzing the set of possible intents defined for the task), and multiple intents detection (how to detect that the user's intent refers to more than one intent and correctly identify each one of them) [7, 8].

Intent recognition is a particularly useful task for conversational systems oriented to FAQ services [1]. In this kind of services, the intent recognizer can be used for both completing the natural language understanding task (by means of the detection of the main information pieces that are present in the user's utterances) and the dialogue management (by means of assigning the user utterance to one of the intents defined as possible responses for the system).

In this paper we present a novel approach for intent detection and classification based on word-embeddings and recurrent neural networks. We have validated our approach with a selection of the corpus acquired with the Hispabot-Covid19 conversational system, which was developed by the Spanish Goverment to provide responses to FAQ related to the pandemics originated by the Covid-19. The results of the evaluation shows the improvement of performance when using named-entities for intent recognition and embeddings adapted to the specific task, compared to our baseline approach (based on raw text with basic pre-processing).

2. Related work

As it has been described in the previous section, user intent detection plays a critical role in question-answering and dialogue systems. Traditional intent detection methods include rule-based template semantic recognition methods and method based on the use of statistical features, such as Naive Bayes, Adaboost, Support Vector Machines, and logistic regression [3, 4].

Current mainstream methods are mainly based on deep learning techniques and the use of word embeddings, which has been probed as a solution to better representational ability and domain extensibility instead of using bag of words [9].

Intent recognition methods based on deep learning techniques can be roughly classified into methods using convolution neural networks [10], recurrent neural networks [11] and their variants (LSTMs and GRUs) [7, 4], the Bidirectional Long short-term Memory (BLSTM) self-attention model [12], the capsule network model [13], the method of joint recognition [14], the use of distances to measure the text similarities (such as TF-IDF) [15], or methods combining several deep learning models [16, 17].

3. The Hispabot-Covid19 dataset

Hispabot-Covid19 is a conversational system developed for the Spanish Government to provide responses to frequently asked questions related to the pandemic originated by the Covid-19 and its implications in Spain. The system received more than 350,000 queries between April and June 2020. The assistant provided information from official sources, such as the Spanish Ministry of Health and the World Health Organization.

A total of 164 intents were defined for the system in order to classify the user's utterances and provide a response associated to each one of them. These intents are related to the symptoms

10.21437/IberSPEECH.2021-10

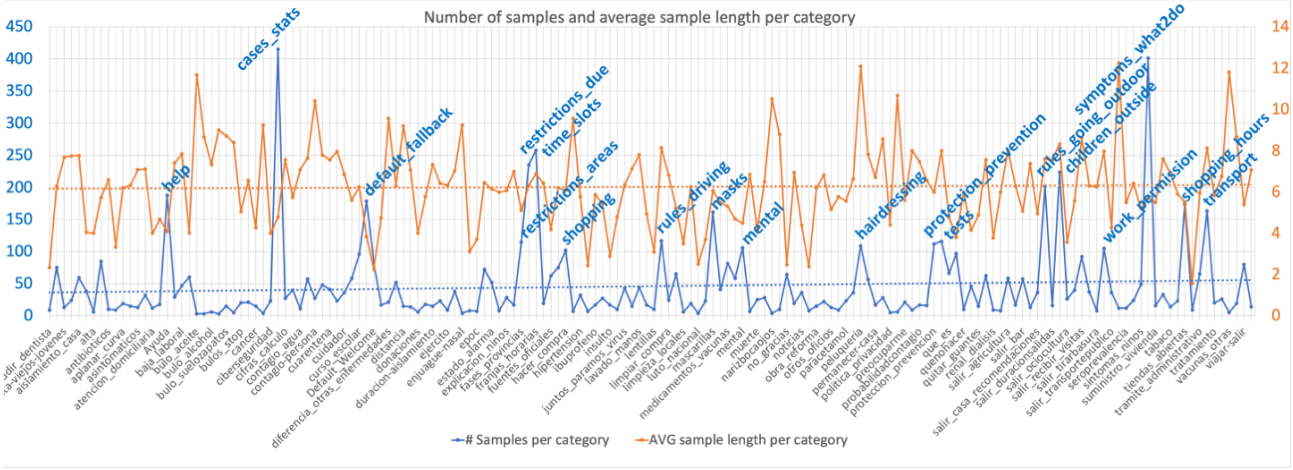


Figure 1: *Histogram of the amount of samples per class and average sample length per class*

of the disease, vulnerable groups, how it is transmitted, how to prevent and protect each self, living with infected people, conditions for isolation, official telephone numbers, among many others. The assistant also incorporated the information published in the Spanish B.O.E. regarding the application of the State of Alarm and the Plan for the Transition towards a New Normality. The assistant did not require or analyze personal data.

All interactions were recorded in the system and processed for continuous retraining and improvement of the chatbot. To this end, new training phrases and new question categories were incorporated daily, misunderstandings have been detected and corrected, and the knowledge base has been restructured according to updates in the information provided. Table 1 summarizes the main features of the selection of the corpus acquired with the Hispabot-Covid19 system that has been used to validate our proposal.

Figure 1 presents a histogram representation with the amount of samples available per each category (in blue) where most frequent categories (above 100 samples) have been highlighted in bold above the corresponding line. Besides, the figure also gives the average sample length per class (in orange). Mean values have also been computed and depicted for both data series (in dotted lines).

4. Model

As a solution for the sentiment analysis we have used a Recurrent Neural Network (RNN), a type of model widely used in the analysis of Twitter messages, for example. RNNs have the ability to process their inputs sequentially, performing the same operation, $h_t = f_W(x_t, h_{t-1})$, on each of the different elements that constitute our input sequence (i.e. words or, to be more exact, their corresponding embeddings), where h_t is the hidden state, t the time step, and W the weights of the network.

Table 1: *Main features of the selection of the Hispabot-Covid19 corpus used in the paper*

Param	Value
Number of samples	7532
Number of intents	164
Number of entities	55
Different words	7658
Average number of words per utterance	6.16
Average number of entities per utterance	0.6
Average number of words per entity	1.22

As it can be observed, the operation is formulated in such a way that the hidden state at each time step depends on the previous hidden states. Hence, the order of the elements in our sequences (i.e. the order of the words) is particularly important. As an immediate consequence, RNNs allow us to handle inputs (i.e. sentences) of variable length, which happens to be an essential feature given the nature of our problem.

Among the different possible architectures of this type of networks, we have opted for the so-called Long Short-Term Memory (LSTM) networks [2], a special type of RNNs that help preventing the typical vanishing gradient problem of standard RNNs by introducing a gating mechanism to ensure proper gradient flow through the network. LSTMs main characteristic is the ability to learn long-term dependencies. To do this, these networks are supported by basic constituent units called *cells* that are provided with mechanisms that allow deciding for each cell what information is preserved from that provided by the previous cells, and what information is provided to the next ones, both depending on the cell's current state.

4.1. Embeddings

(Word) embeddings are vector-type representations obtained for words in reduced-dimensional vector spaces where semantically similar words are always close to each other. The Fasttext project [18], recently open-sourced by Facebook Research, enables a fast and effective method to learn word embeddings that are very useful in text classification, clustering and information retrieval. In this work, the proposed model uses Fasttext word embeddings to represent the vectors for the words as input of the network.

At the time of training, FastText trains by sliding a window over the input text and either learning the target word from the remaining context (also known as continuous bag of words, CBOW), or all the context words from the target word ("Skip-gram"). Learning can be viewed as a series of updates to a neural network with two layers of weights and three layers of neurons, in which the outer layer has one neuron for each word

in the vocabulary and the hidden layer has as many neurons as there are dimensions in the embedding space. This approach is very similar to Word2Vec [19]. However, unlike Word2Vec, fastText might also learn vectors for sub-parts of words: so-called character n-grams. This ensures that for instance the words love, loved and beloved all have similar vector representations, even if they tend to show up in different contexts. This feature enhances learning on heavily inflected languages [20].

4.2. Model description

Our approach is based on a 2-layer Bidirectional-LSTM model with a deep self-attention mechanism which is represented in Figure 2. The model is implemented in Pytorch [21] and based on the architecture proposed in [22].

4.2.1. Embedding layer

The model is designed to work with sequences of words as inputs, thus allowing us to process any type of sentence. For this, a first embedding layer is provided that collects the embeddings $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ corresponding to each of the words $w_1, w_2, \ldots, w_N$ constituing the sentence we want to process, where N is the number of words in our sentence. We initialize the weights of the embedding layer with our pre-trained word embeddings.

4.2.2. Bi-LSTM layer

A standard LSTM model behaves in a unidirectional way, that is, the network takes as input the direct sequence of word embeddings and produces the outputs $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N$, where $\boldsymbol{h}_i$ is the hidden state of the LSTM cell at time step i, summarizing all the information that the network has accumulated from our sentence up to word w_i.

Instead, we have used a *bi-directional LSTM* (Bi-LSTM) that allows us to collect such information in both directions. In particular, a Bi-LSTM consists of 2 LSTMs, a *forward* $\overrightarrow{LSTM}$ that allows the analysis of the sentence from w_1 to w_N, and an *inverse or backward* $\overleftarrow{LSTM}$ which allows a similar analysis to be carried out but in the opposite direction, from w_N to w_1. To obtain the definitive outputs of our Bi-LSTM layer, we simply concatenate for each word the outputs obtained from the analysis performed in each specific direction (see Equation 1 in which $||$ corresponds to the concatenation operator and L to the size of each LSTM).

$$h_i = \overrightarrow{h_i} || \overleftarrow{h_i} \text{ , where } h_i \in R^{2L} \tag{1}$$

4.2.3. Attention layer

In order to identify the most informative words when determining the polarity of the sentence, the model uses a deep self-attention mechanism. Thus, actual importance and contribution of each word is estimated by means of a multilayer perceptron (MLP) composed of 2 layers with a non-linear activation function (*tanh*) similar to that proposed in [23].

The MLP learns the attention function g as a probability distribution on the hidden states h_i, that allows us to obtain the attention weights a_i that each word receives. As the output of the attention layer the model simply computes the convex combination r of the LSTM outputs h_i with weights a_i, where a convex combination is a linear combination of points where all the coefficients are non-negative and add up to 1.

4.2.4. Output layer

Finally, we use r as a feature vector which we feed to a final task-specific layer for classification. In particular, we use a fully-connected layer, followed by a *softmax* operation, which outputs the probability distribution over the classes.

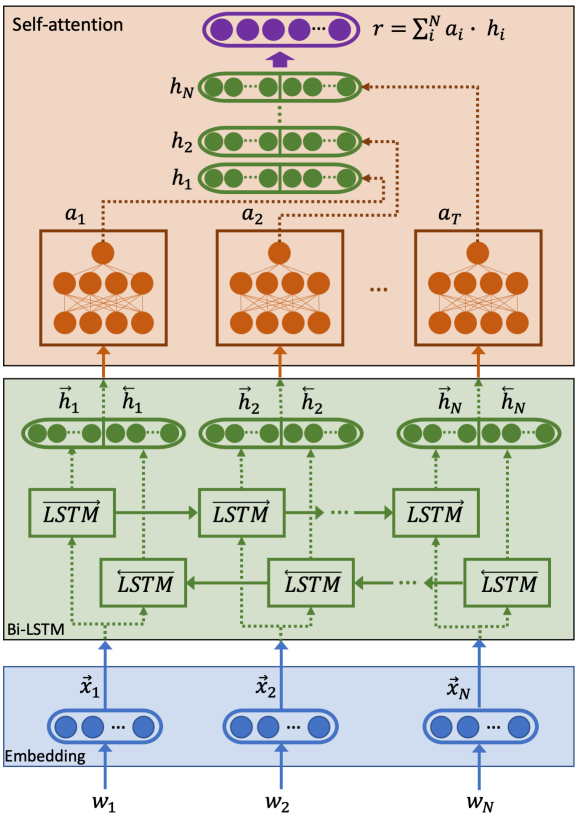


Figure 2: *Proposed model.*

5. Evaluation

According to the proposed methodology, we first pre-process the input text to transform it as a list of words. For this purpose, special characters are removed and numerical data are replaced by tokens, among other transformations. Subsequently, the remaining words are then mapped and transformed into their corresponding embeddings as the network can only work with numeric representations of each word. The type of word embeddings used in this work is described later in subsection 5.4. Specifically, the two different sets of word embeddings that have been applied are detailed in the next subsection. Once the input text is transformed into word embeddings by the Embeddings layer, the network is used to finally classify the text according to the identifed classes of the problem at hand.

5.1. Experimental setup

To prevent overfitting all the experiments have been carried out following a 5-fold cross-validation scheme. Each setup has been trained for 100 epochs and 'Stop-Early' has been adopted as the stopping criterion. Statistical tests for signifcance have been also applied to the results obtained in order to evaluate if there is a significant diference of performance depending on the adopted confguration.

Table 2: *Summary of results*

Model	Inputs	Embedding type	F1 score (%)
1	Pre-processed	General pre-trained	67.72
2	NER based	General pre-trained	72.58
3	NER based	Domain-specific	75.33

Table 3: *Length analysis*

	Total	Correct	Wrong
Number of samples	7532	5467	2065
Average length (in words)	6.16	5.87	6.93
Percentage	100	72.58	27.42

Although it could be possible to evaluate if there is any significant performance improvement when adjusting the hyperparameters of our model, in this work we did not aim at fine-tuning the network to achieve better performance than any other model in the state of the art, but rather to evaluate the sensitivity of our network in response to different pre-processing and embeddings choices. Another important aim has been to also develop a baseline for a deeper understanding of the dataset when addressing the problem of intent recognition.

The model was trained during 100 epochs using an Adam optimizer [24], with initial learning rate of 0.001, batch size of 32, and early-stopping after 5 epochs without improvement in the F1 classification score. Both bi-LSTM and attention layers had a 0.3 dropout rate. The encoder layers had a size L of 100. As a way to increase input variability from epoch to epoch, input embeddings are randomly added white noise with 0.15 probability rate in order to increase the robustness of the model.

As it can be deduced from Figure 1, some classes have more training examples than others. Hence, to prevent introducing bias in our models we apply class weights to the loss function, penalizing more the misclassification of under-represented classes. These weights are computed as the inverse frequencies of the classes in the training set.

5.2. Preprocessing of training data

Data preprocessing of our dataset goes through four main steps. These steps, described hereafter, are mainly intended to remove the noise present in the sentences and to encode the sentences in a way that can be used in the training of our network.

- All words are written in lowercase letters.
- Special words including emails, percentages, money, phone numbers, times, dates, urls and/or hashtags are assimilated and replaced by a special tokens, such as MAIL, DATE, URL,... to prevent information from being lost during data representation.
- The words contained in the datasets (or their corresponding entities) are replaced by their respective pre-trained vectors.
- The words, which are present in the datasets but not in the pre-trained word vectors, are considered Out Of Vocabulary (OOVs) words and simply discarded.

5.3. Named entity recognition

Named-entity recognition (NER) (also known as entity identification, entity chunking and entity extraction) is a subtask of information extraction that seeks to locate and classify elements in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. [25]. In this work, we have not implemented a Named Entity Recognizer but simply relied on manually annotated entities. All the sequences of words in the sentences of the dataset that are linked to relevant entities for the intent recognition task, such as dates, activities, symptoms, or covid and location names, have been carefully identified and annotated. Hence, the dataset comes with full annotation of entity-related items that will be useful for engineering feature extractors for Named Entity Recognition in the future.

5.4. Word embeddings

Using pre-trained word embeddings vectors as inputs of machine learning models is always a interesting option when training data corpora are restricted or limited. Hence, for our first approach we have used a set of pre-trained Fasttext vectors made up of a total of 2 million embeddings of dimension $W = 300$ generated for the Spanish language from different texts massively and automatically retrieved from the web and, in particular, from Wikipedia [26].

Alternatively, we have also used Fasttext to train task-specific embeddings from scratch. A setup similar to the one used for general pre-trained models was adopted. As a result, we also obtained CBOW models but with a smaller size: dimension was set to 100 in this case, consistently with the smaller size of our dataset.

5.5. Results

We have evaluated three different models whose main characteristics and results are detailed in Table 2. For the calculation of F1 we used the *weighted* version that takes into account the number of examples available for each different class.

The first model, that was adopted as our baseline, was directly trained from the available sentences after applying them the basic pre-processing described in section 5.2. General pre-trained embeddings were used for the tokenized words.

Compared to our baseline, the second model successfully introduced the use of the named-entities version of our dataset while the third one, the top performing model, also combined it together with task-specific embeddings trained on our dataset.

As it can be observed in Figure 1, there are some evident asymmetries in the distribution of the data. Therefore, we decided to further explore whether the sentence length affects the final performance on the assumption that shorter sentences are more difficult to classify. To that end, we computed the average length of correctly and incorrectly recognized sentences, respectively. However, the results obtained of this length analysis, summarized in Table 3 for the second model, suggested that the hypothesis did not hold for our data.

Finally, although omitted in this work, error analysis indicates that errors are mainly due to semantically overlapped categories, such as “shopping” and “opening hours” or “traveling” and “transports”, which suggests that a simplified and reduced set of intent categories may be also considered.

6. Conclusions

In this paper, the intent recognition results obtained in the evaluation of different pre-processing and embeddings configura-

tions of an attentive recurrent network are presented. The results indicate that there may be a large improvement of performance when using named-entities for intent recognition, relying on manual annotation, compared to our baseline approach (based on raw text with basic pre-processing), which suggests that conveniently parsed text should more accurately link entities to intents. A similar result may also be observed when relying on embeddings adapted to the specific task (i.e. trained on task-specific data), which suggests that, although general-purpose embeddings pre-trained on larger corpora reasonably fit the task, training corpus size is suitable for optimizing and building custom vectors for these specific domain and task. Finally, the Hispabot-Covid19 data set is a novel resource for researchers. This is the very first work addressing intent recognition based upon this dataset, thus all the experiments carried out in this paper should provide a relevant insight to those researchers.

7. Acknowledgements

The work leading to these results has been supported by the Spanish Ministry of Economy, Industry and Competitiveness through CAVIAR (MINECO, TEC2017-84593-C2-1-R) and AMIC (MINECO, TIN2017-85854-C4-4-R) projects (AEI/FEDER, UE).

8. References

[1] M. McTear, *Conversational AI. Dialogue systems, Conversational Agents, and Chatbots*. Morgan and Claypool Publishers, 2020.

[2] M. McTear, Z. Callejas, and D. Griol, *The Conversational Interface: Talking to Smart Devices*. Springer, 2016.

[3] J. Liu, Y. Li, and M. Lin, "Review of intent detection methods in the human-machine dialogue system," *Journal of Physics*, vol. 1267, 2019.

[4] S. Ravuri and A. Stoicke, "A comparative study of neural network models for lexical intent classification," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU-2015)*, 2015, pp. 702–709.

[5] S. Larson, A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzano, L. Tang, and J. Mars, "An evaluation dataset for intent classification and out-of-scope prediction," in *Proc. of EMNLP-IJCNLP'19*, 2019.

[6] W. Xie, D. Gao, R. Ding, and T. Hao, "A feature-enriched method for user intent classification by leveraging semantic tag expansion," *LNAI*, vol. 11109, pp. 224–234.

[7] G. H. E. A. e. a. Firdaus, M., "A deep multi-task model for dialogue act classification, intent detection and slot filling," *Cognitive Computation*, 2020.

[8] U. Park, J.-S. Kang, V. Gonuguntla, K. C. Veluvolu, and M. Lee, "Human implicit intent recognition based on the phase synchrony of eeg signals," *Pattern Recognition Letters*, vol. 66, pp. 144–152.

[9] J. Kim, G. Tur, and A. C. et al., "Intent detection using semantically enriched word embeddings," in *Proc. of IEEE Workshop on Spoken Language Technology*, Brussels, Belgium, 2016, pp. 414–419.

[10] H. Hashemi, A. Asiaee, and R. Kraft, "Query intent detection using convolutional neural networks," in *Proc. of Int. Conference on Web Search and Data Mining, Workshop on Query Understanding*, 2016.

[11] A. Bhargava, A. Celikyilmaz, and D. H. et al., "Easy contextual intent prediction and slot detection," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 8337–8341.

[12] K. Sreelakshmi, P. Rafeeque, S. Sreetha, and E. Gayathri, "Deep bi-directional lstm network for query intent detection," in *Proc. of 8th International Conference on Advances in Computing Communications (ICACC-2018)*, 2018, pp. 8337–8341.

[13] C. Xia, C. Zhang, X. Yan, Y. Chang, and P. Yu, "Zero-shot user intent detection via capsule neural networks," in *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018, pp. 3090–3099.

[14] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," in *Proc. of 17th Annual Conference of the International Speech Communication Association*, 2016, pp. 685–689.

[15] R. Rojowiec, M. C. Fink, and B. Roth, "Intent recognition in doctor-patient interviews," in *Proc. of 12th Conference on Language Resources and Evaluation (LREC-2020)*, 2020, pp. 702–709.

[16] H. Yu, X. Feng, and L. L. et al., "Identification method of user's medical intent in chatting robot," *Journal of Computer Applications*, vol. 38, no. 8, pp. 2170–2174, 2020.

[17] C. Yang and C. Feng, "Multi-intent recognition model with combination of syntactic feature and convolution neural network," *Journal of Computer Applications*, vol. 38, no. 7, pp. 1839–1845, 2018.

[18] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[20] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017. [Online]. Available: https://www.aclweb.org/anthology/Q17-1010

[21] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS 2017 Workshop on Autodiff*, 2017. [Online]. Available: https://openreview.net/forum?id=BJJsrmfCZ

[22] C. Baziotis, A. Nikolaos, A. Chronopoulou, A. Kolovou, G. Paraskevopoulos, N. Ellinas, S. Narayanan, and A. Potamianos, "Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning," *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018. [Online]. Available: http://dx.doi.org/10.18653/v1/S18-1037

[23] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, "Deep learning for user comment moderation," *Proceedings of the First Workshop on Abusive Language Online*, 2017. [Online]. Available: http://dx.doi.org/10.18653/v1/W17-3004

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[25] U. Yasavur, J. Travieso, C. Lisetti, and N. Rishe, "Sentiment analysis using dependency trees and named-entities," in *FLAIRS Conference*, 2014.

[26] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Contrasting the Emotions identified in Spanish TV debates and in Human-Machine Interactions

Mikel deVelasco, Raquel Justo, Leila Ben Letaifa, M. Inés Torres

Speech Interactve Research Group, Universidad del País Vasco UPV/EHU
{mikel.develasco, raquel.justo, leila.benletaifa, manes.torres}@ehu.es

Abstract

This work is aimed to contrast the similarities and differences for the emotions identified in two very different scenarios: human-to-human interaction on Spanish TV debates and human-machine interaction with a virtual agent in Spanish. To this end we developed a crowd annotation procedure to label the speech signal in terms of both, emotional categories and Valence-Arousal-Dominance models. The analysis of these data showed interesting findings that allowed to profile both the speakers and the task. Then, Convolutional Neural Networks were used for the automatic classification of the emotional samples in both tasks. Experimental results drew up a different human behavior in both tasks and outlined different speaker profiles.

Index Terms: emotions recognition from speech, perception, communication, human-machine interaction, crowd annotation, speech processing.

1. Introduction

Speech signal includes information about the personal characteristics of the speaker, the content of the message delivered or the language used to code it, among others [1]. The analysis of the speech also allows to estimate, to some extent, the current emotional status of the speaker [2, 3, 4], even the basal mood, or the probability to be suffering a particular mental disease [5]. However, speech may also be influenced by several other variables, such as the habits of the speaker, his personality, culture or the particular task being performed [6, 7]

This work is aimed to contrast the similarities and differences for the emotions identified in two very different scenarios: human-to-human interaction on Spanish TV debates and human-machine interaction with a virtual agent in Spanish. Thus, we focus on spontaneous emotions appearing in each task that show significant differences to the six basic emotions [8] that have been many times simulated by professional actors [9, 10, 11] and recorded in the lab [12]. In fact, spontaneous emotions have been hypothesized to be extremely task dependent [2, 3, 7, 4, 6]. Further to this, emotions cannot be unambiguously identified. As a consequence, not even expert labelling procedure can lead to a ground truth for learning. As an alternative, crowd annotation implementing perception experiments has also been proposed as a way to establish the ground truth [13]. However, human perception of emotions does not usually show a high agreement. As a consequence, a certain ambiguity and uncertainty always remains, which adds an stochastic component to the emotion identification problem.

In order to verify whether actually the task plays a significant role when dealing with emotion detection, a preliminary comparison of the emotional content in two very different Spanish tasks was carried out in this research work. To this end, we chose the following set of features to be analysed: agreement in crowd annotation, perceived emotions and significance in the particular task, distribution of categories in both tasks, distribution of dimensional axes of emotions, namely Valence, Arousal and Dominance (VAD), and the representation of the categories into the 3D VAD model. An additional contribution is the comparative analysis of the results in terms of categories of the automatic classification of the samples based on Convolutional Neural Networks (CNN).

Section 2 describes the two tasks addressed as well as the annotation procedure and its outcomes. Then Section 3 develops the analysis of emotional content of the corpora and Section 4 describes the preliminary classification experiments carried out. Finally, Section 5 summarizes the concluding remarks and future work.

2. Perception of emotions

2.1. Description of the tasks

TV Debates Firstly, a data-set that gathers real human-human conversations extracted from TV debates, specifically the Spanish TV program "La 6 Noche", was selected. In this weekly broadcasted show, hot news of the week are addressed by using social and political debate panels that were led by two moderators. There is a very wide range of talk-show guests (politicians, journalists, etc.) who analyse, from their perspective, social topics. Given that the topics under discussion are usually controversial it is expected to have emotionally rich interactions. However, the participants are used to speak in public so they do not lose control of the situation and even if they might overreact sometimes, it is a real scenario, when emotions are subtle. The spontaneity in this situation makes a great difference from scenarios with acted emotions as shown in [2]. The selected programs were broadcasted during the electoral campaign of the Spanish general elections in December 2015.

Elder interaction with simulated virtual agent Empathic is a European Research & Innovation Project [1] [14, 7] that implements personalized virtual coaching interactions to promote healthy and independent aging. As a part of the project, a series of spontaneous conversations between elderly and a Wizard of OZ (WOZ) have been recorded in three languages: Spanish, French and Norwegian. WOZ's technique allows users to believe that they are communicating with a human (and not a machine) in order to make their reaction more natural [7]. The conversations are related to four main topics: leisure, nutrition, physical activity and social and family relationships [14, 7]. In this work we focused on the Spanish dialogues that were recorded by 79 speakers resulting in 7 hours and 15 minutes of audio extracted from the recordings [3].

[1] www.empathic-project.eu

10.21437/IberSPEECH.2021-11

2.2. Crowd perception

TV Debates and Spanish Virtual agent interaction data were labeled in terms of emotions using the crowd annotation technique. To begin with, we automatically extracted segments of audio that we estimated to match a clause. A clause can be defined as "a sequence of words grouped together on semantic or functional basis" [15]. Thus, we can hypothesize that the emotional status does not change inside a clause. This procedures allowed to get 4118 chunks from the TV Debate corpus and 2000 from the Virtual agent corpus. Then, all these segments were crowd annotated by native speakers. To this end, both categorical and VAD model of emotions were considered. For the categorical model we first consider the categories proposed in [16] and then we reduce and adapt the list of each of the tasks. For TV Debates task we selected a list of ten labels to be considered by annotators. Then we added three questions to annotate the perception of each of the axes of the dimensional model, namely Valence, Arousal and Dominance

Three of them are related to the arousal: Excited, Slightly excited and Neutral.Valence is annotated as Positive or Slightly positive or Neutral or Slightly negative or Negative in TV Debates. For Virtual Agent task, valence is assigned one of only three labels that are Positive, Neutral and Negative for valence. The dominance labels are: Rather dominant / controlling the situation, Rather intimidated / defensive, and Neither dominant nor intimidated.. The whole questionnaire is reported in [2]. For Virtual Agent task we also selected list of ten categories adapted to the task that differs from the previous one. As an example *Sad* was only included in this task whereas *Annoyed* was only proposed to TV Debates annotators.

Annotators agreement Each audio segment was annotated by 5 different annotators. Table 1 shows the statistics of agreement per audio chunk for the categorical model. This table shows that for about 70% of the data and in both tasks, the agreement is 3/5 or 2/5. This confirms the ambiguity and subjectivity of the task. Moreover the Krippendorff's $alpha$ coefficient was also low for both tasks resulting in 0.11 and 0.13 values respectively. This coefficient reflects the agreement degree but is very dependent on the number of labels, which was high and sometimes difficult to be perceived.

In the rest of the document, we do not consider samples with agreement below 0.6, which means we have used the 64.13% of the corpus for the TV debates task and the 66.20% of the Virtual Agent task.

Table 1: *Statistics of the agreement per audio chunk*

	TV Debates		**Virtual Agent**	
Agr	No. audios	% audios	No. audios	%. audios
5/5	197	4.72%	149	7.45%
4/5	799	19.40%	421	21.05%
3/5	1645	39.95%	754	37.7%
2/5	1431	34.75%	636	31.8%
1/5	46	1.18%	40	2%
Tot. audios	4118		2000	

Annotation labels The defined sets of labels were then reduced by merging overlapping categories that we selected for the tag pairs with high level of confusion among them in the annotation procedure. Then, a minimum agreement of 0.6 (3/5) was requested for each sample as a well as a minimum number of samples. Table 2 shows the resulting list of categories considered for each task along with the percentage of samples. This Table shows that different categories appear in each corpus. Some of them could be equivalent, such as *Calm/Indiferent* and *Calm/relaxed* but *annoyed/tense* does not appear in Virtual Agent task whereas puzzled is not in the list for TV debates.

Table 2 also shows that both data-sets are imbalanced, being the *Calm* category the majority class with around 75% of the samples. This reflects the spontaneous nature of the data. There are more positive emotions in the Virtual Agent annotations and more negative emotions in TV Debates. This difference comes from the tasks characteristics. During political debates, people try to convince or even impose their opinions on other interlocutors. However, during the coaching sessions, people speak with a machine. They are quiet and paying attention to the answers to their expectations.

For the dimensional model we got a set of scale values for each axe. For for Arousal we proposed Neutral, Slightly excited and Excited in both databases. For Dominance we proposed Rather intimidated / defensive, Neither dominant or intimidated, Rather dominant / controlling the situation fro both databases. For Valence we got Negative, Slightly Negative, Neither negative or positive, Slightly Positive and Positive for TV Debates whereas we reduced the scale to Rather Negative, Neither negative or positive, Rather Positive for the Virtual Agent task. These dimensions are considered in Section 3

Table 2: *Categories more frequent in the corpora*

TV Debates		**Virtual Agent**	
Category	% audios	Category	% audios
Calm/Indiferent	73.64	Calm/Relaxed.	78.32
Annoyed/Tense	14.32	Happy/Pleased	8.76
Enthusiast	4.72	Interested	5.66
Satisfied	3.23	Puzzled	2.95
Worried	2.12		
Interested.	1.57		
Others	0.40	Others	4.31

3. Analysis of emotions

Figure 1 shows the probability density function of each variable (Valence, Arousal, Dominance) of VAD model that has been obtained by a Gaussian kernel density estimator (upper row). Figure 1 also shows different 2D projections of sample distribution in the 3D space (row below), representing each scenario in a different colour. When regarding Arousal, Virtual Agent seems to work in a very neutral scenario where excitement is almost absent. In TV debates, although neutrality is also predominant, some excitement is perceived, due to the debate nature of the conversations. Valence distribution shows a clear deviation towards positive values when considering Virtual Agent scenario, a sign of the good acceptance of the system among the users, whereas in TV debates neutrality is predominant with only a slight nuance towards positiveness. On the contrary Dominance is shifted towards Dominant values, in TV debates, but keeps

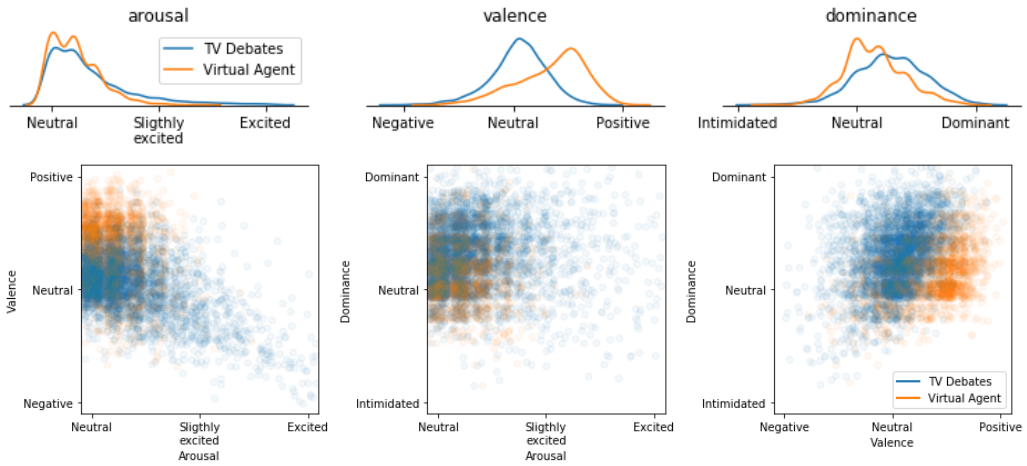


Figure 1: *VAD representation.*

neutral when users interact with the Virtual Agent. These results correlate well with the kind of audios we are dealing with in the two scenarios. In TV debates, people express themselves without getting angry (low levels of excitement) but in a very assertive way (quite high dominance levels). Additionally they appear to be neutral when communicating their opinions (valence tends to be neutral or slightly positive). In the Virtual Agent scenario the users are volunteers, with a good predisposition, and thus they seem to be pleased with the system (Positive Valence values). They are relaxed talking to the agent (levels of excitement tend to neutrality) and although they do not have to convince anyone they know well what they are talking about and are not intimidated (dominance values are around neutrality with a slight shift to the right).

The categorical model is also considered in this work and each category is represented in the 3D VAD space for comparison purposes. Specifically, the average of the Valence, Arousal and Dominance values of all the audios labeled within a specific category was computed and the resulting value was represented as a point in the 3D space. Figure 2 shows 2D projection of the resulting representation. If we focus on TV Debates, it can be noticed that *Interested* and *Worried*, the least representative categories, according to Table 2, are very close to the category with the highest number of samples, *Calm/Indifferent*, in all the 2D projections (purple, orange and deep blue points), so they were merged in an only one category. The same happens with Enthusiastic and Satisfied (light blue and green points). When considering Virtual Agent scenario although the category *Calm/Relaxed* is the most relevant one with more than the 75% of the samples we decided to keep the remaining categories because the fusion is not as clear as in the previous case, as shown in Figure 2. Thus the final set of categories used for the classification experiments reported in this work (Sectiion **??**) is the following one for TV Debates: 1) *Annoyed/Tense*, 2) *Enthusiastic + Satisfied*, 3) *Calm/Indifferent + Interested + Worried* and for as the Virtual Agent the list is: 1) *Calm/Relaxed*, 2) *Happy/Pleased*, 3) *Interested*, 4) *Puzzled*.

As shown above, there are some categories that are not in both sets due to the nature of the different tasks, like Annoyed/Tense that is only In TV Debates or Puzzled that only appears in the interaction with the Virtual Agent. Moreover, Figure 2 shows that there is not any point in Virtual Agent scenario around the location of the red point (Annoyed/Tense) of TV Debates (higher excitement levels and negative values of Valence), which is in fact quite separated from the other categories. The same happens with *Puzzled* represented by the brown point (low levels of Valence and Dominance) that has not any representation in TV Debates and it is a bit separated from the other categories in Virtual Agent scenario. This correlates well with the idea that people interacting with the Virtual Agent are not in general annoyed or tense, while this is a quite common feeling in a debate. Furthermore, speakers in the debates do not usually show that they are in an unexpected situation, since it can be interpreted as a weak point, while it is quite easy to imagine it in the interaction with a machine. There are also categories, like *Calm* that has a similar location in both scenarios but with higher values of Valence for Virtual Agent interactions. That is, the users interacting with the Virtual Agent perceived as calm tend to be more positive than the ones in TV Debates. The same happens with *Enthusiastic + Satisfied* from TV Debates and *Happy/Pleased* from Virtual Agent, that although they are very close in their location in both scenarios (with a very similar meaning) *Happy/Pleased* seems to have more positive Valence values than *Enthusiastic + Satisfied*, but a bit lower Dominance and Arousal values.

4. Experiments and results

To complete the work, some classification problems were carried out in both tasks described in Section 2.1. For TV Debates, 4118 chunks were selected distributed in the 3 classes mentioned avobe (*Annoyed/Tense*, *Enthusiastic + Satisfied*, and *Calm/Indifferent + Interested + Worried*) and for the Virtual Agent, 2000 samples were selected divided into 4 classes (*Calm/Relaxed*, *Happy/Pleased*, *Interested*, and *Puzzled*).

One of the challenges of both data-sets is the different length of each audio sample. Some kind of Neural Networks are specifically well suited to deal with this problem and given that deep learning is the state of the art in many AI areas, including emotion recognition, a Convolutional Neural Network architecture was designed for this work. Let us note that in [17] a neural network architecture provided promising results when comparing ot to classical Support Vector Machines, for a regression problem over the task related to TV debates.

The number of samples in both data-sets are also a challenge. It makes nonsense to try to identify the emotions from

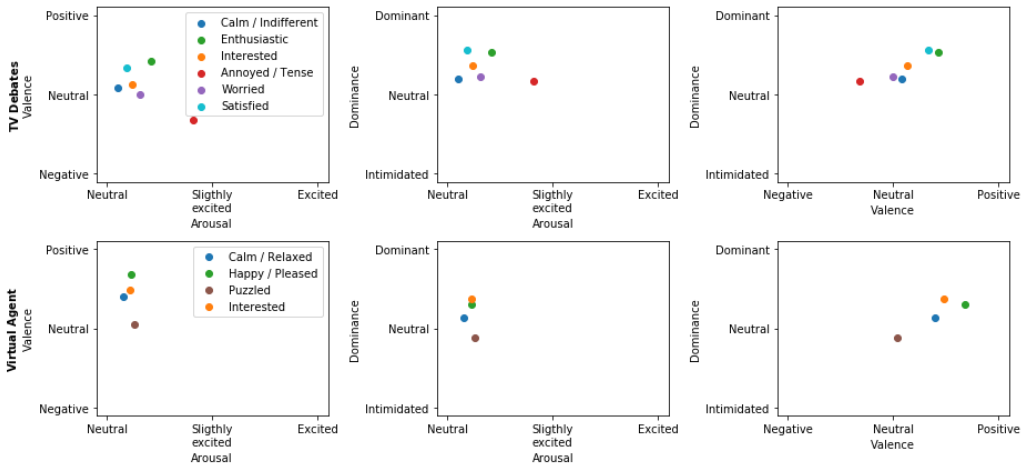


Figure 2: *Categories in dimensional representation.*

raw-audio. Different works suggest that there is not a standard audio feature-set that works well for all emotion recognition corpora [18, 19, 20]. In this context, we decided to use the audio Mel-frequency spectrogram as the classifier's input. It is known that the spectrogram encodes almost all audio information and should be possible to identify from that.

Figure 3 shows the architecture of the network used in this work. It takes the mel-spectrogram input and reduce both mel-frequency and time dimensions using 2D convolutions and max-poolings (red boxes). This sub-network reduces time dimension but creates richer audio representation. Then, the network takes the new representation and try to classify each time step. After classifying all time steps, the network averages it in order to provide an output for the input audio.

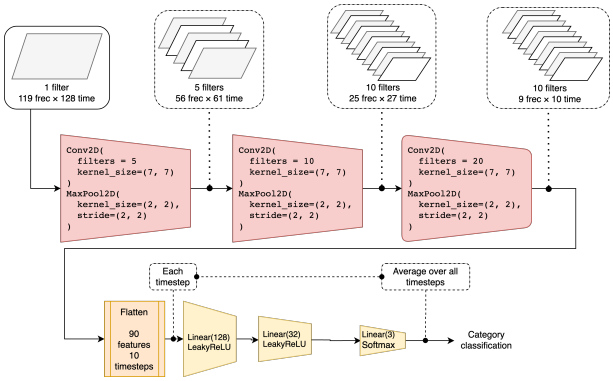


Figure 3: *Architecture of the Network used*

In the training process, several decisions were chosen. On the one hand, the network will only see a sub-part of the full audio. Thus, the training process is easier if all the batches work with the same input length, which can be considered as a dropout mechanism. On the other hand, an repetition over-sampling method was chosen, where all the non-majority class samples were provided 5 times. It helps the network to avoid the exclusive prediction of the majority class. Adam optimizer is used with a learning rate of 0.001 and 150 epochs were training on all database. These experiments were carried out over a 10-fold cross-validation procedure.

Classification results are given in Table 3. Most promising results come from TV Debates, in fact, the model guesses 72% of the test samples, and achieves a F1 Score of 0.59, that can be considered a good result taking into account the ambiguity and subjectivity of the task.

As expected, the category *Calm/Indifferent + Interested + Worried* got better results since it is the majority class with a F1 Score of 0.82.In contrast, *Annoyed/Tense* and *Enthusiastic + Satisfied* perform a little bit worse, with a 0.56 and 0.43 in F1 Score.

Table 3: *Evaluation of the classification results for the categorical model*

TV Debates				**Virtual Agent**			
Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
72%	0.56	0.66	0.59	74%	0.32	0.27	0.27

Nevertheless, Virtual Agent experiments obtained lower results. Table 3 show a very high accuracy (74% of the samples) along with low values of F1, precision and recall values. The majority class, i.e.*Calm*, achieved an F1 score of 0.88 whereas all minority classes remain under 0.32 fro F1. This is mainly due to the huge imbalance of this data-set along with the very reduced number of samples.

5. Conclusions

This work provides a comparison of the emotional content in two different Spanish corpora dealing with very different tasks. The emotional labels, associated to spontaneous emotions, were achieved by means of perception experiments using crowd annotation. The agreement among the annotators was considered to build the ground truth. The analysis carried out shows the main differences associated to each task, in terms of both, the emotional category distribution and the level of Valence, Arousal and Dominance and brings out the relevance of the task when addressing an emotion recognition problems. This analysis also highlights that the perception experiments carried out were able to outline a different speaker profile for each of the tasks. Thus, crowd annotation seems to be valid approach for emotions. Finally, some preliminary classification experiments

were also conducted showing very promising results for TV Debate task whereas the Virtual Agent task needs more samples and a more sophisticated oversampling method. Future work includes a deeper and interrelated analysis of the data as well getting a higher number of annotated samples for the Virtual Agent classification task.

6. Acknowledgements

The research presented in this paper is conducted as part of the AMIC and EMPATHIC projects project that have received funding from the Spanish Minister of Science under grant TIN2017-85854-C4-3-R and from the European Union's Horizon 2020 research and innovation program under grant agreement No 769872. First author has also received a PhD scholarship from the University of the Basque Country UPV/EHU, PIF17/310.

7. References

[1] A. López-Zorrilla, N. Dugan, M. Torres, C. Glackin, G. Chollet, and N. Cannings, "Some asr experiments using deep neural networks on spanish databases," in *IberSpeech*, Lisbon, 2016, pp. 149–158.

[2] M. deVelasco, R. Justo, A. López-Zorrilla, and M. Torres, "Can spontaneous emotions be detected from speech on tv political debates?" in *Proceedings of the 10th IEEE International Conference on Cognitive Infocommunications*, Naples, 2019.

[3] L. B. Letaifa, M. I. Torres, and R. Justo, "Adding dimensional features for emotion recognition on speech," in *IEEE International Conference on Advanced Technologies for Signal and Image Processing*, Tunisia, 2020.

[4] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Interspeech*, 2018.

[5] E. L. Campbell, L. Docío-Fernández, J. J. Raboso, and C. García-Mateo, "Alzheimer's dementia detection from audio and text modalities," 2020.

[6] B. Schuller, F. Weninger, Y. Zhang, F. Ringeval, A. Batliner, S. Steidl, F. Eyben, E. Marchi, A. Vinciarelli, K. Scherer, M. Chetouani, and M. Mortillaro, "Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge," *Computer Speech Language*, vol. 53, pp. 156 – 180, 2019.

[7] R. Justo, L. B. Letaifa, C. Palmero, E. G. Fraile, A. Johansen, A. Vazquez, G. Cordasco, S. Schlogl, B. F. Ruanova, M. Silva, S. Escalera, M. D. Velasco, J. T. Laranga, A. Esposito, M. Kornes, and M. I. Torres, "Analysis of the interaction between elderly people and a simulated virtual coach," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, pp. 6125–6140, 2020.

[8] P. A. Davidson. R. J., Ekman, *Nature of emotion: Fundamental questions*, ser. Oxford University Press, P. E. . R. J. Davidson, Ed. New York: Springer, 1994.

[9] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions, IJSE*, pp. 68–99, 2010.

[10] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062 – 1087, 2011, sensing Emotion and Affect - Facing Realism in Speech Processing.

[11] S. E. Eskimez, K. Imade, N. Yang, M. Sturge-Apple, Z. Duan, and W. Heinzelman, "Emotion classification: How does an automated system compare to naive human coders?" in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2274–2278.

[12] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition."

[13] J. Sager, R. Shankar, J. Reinhold, and A. Venkataraman, "VESUS: A Crowd-Annotated Database to Study Emotion Production and Perception in Spoken English," in *Proc. Interspeech 2019*, 2019, pp. 316–320. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1413

[14] M. I. Torres, J. M. Olaso, C. Montenegro, R. Santana, A. Vázquez, R. Justo, J. A. Lozano, S. Schlögl, G. Chollet, N. Dugan, M. Irvine, N. Glackin, C. Pickard, A. Esposito, G. Cordasco, A. Troncone, D. Petrovska-Delacretaz, A. Mtibaa, M. A. Hmani, M. S. Korsnes, L. J. Martinussen, S. Escalera, C. P. Cantariño, O. Deroo, O. Gordeeva, J. Tenorio-Laranga, E. Gonzalez-Fraile, B. Fernandez-Ruanova, and A. Gonzalez-Pinto, "The empathic project: Mid-term achievements," in *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, ser. PETRA '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 629–638.

[15] A. Esposito, V. Stejskal, and Z. Smékal, "Cognitive role of speech pauses and algorithmic considerations for their processing," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 22, pp. 1073–1088, 2008.

[16] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proceedings of the National Academy of Sciences*, vol. 114, pp. E7900 – E7909, 2017.

[17] M. de Velasco, R. Justo, J. Antón, M. Carrilero, and M. I. Torres, "Emotion detection from speech and text," in *Fourth International Conference, IberSPEECH 2018, Barcelona, Spain, 21-23 November 2018, Proceedings*, J. Luque, A. Bonafonte, F. A. Pujol, and A. J. S. Teixeira, Eds. ISCA, 2018, pp. 68–71. [Online]. Available: https://doi.org/10.21437/IberSPEECH.2018-15

[18] L. Tian, J. D. Moore, and C. Lai, "Emotion recognition in spontaneous and acted dialogues," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 698–704.

[19] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," 2017.

[20] S. Parthasarathy and I. Tashev, "Convolutional neural network techniques for speech emotion recognition," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 121–125.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

A proposal for emotion recognition using speech features, transfer learning and convolutional neural networks

Roberto Móstoles[1], *David Griol*[2], *Zoraida Callejas*[2], *Fernando Fernández-Martínez*[3]

[1] Universidad Carlos III de Madrid. Madrid (Spain)
[2] Dept. of Languages and Computer Systems, University of Granada. Granada (Spain)
[3]Speech Technology Group, Center for Information Processing and Telecommunications, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid, Spain

100346034@alumnos.uc3m.es, dgriol@ugr.es, zoraida@ugr.es, fernando.fernandezm@upm.es

Abstract

In this paper, we present a proposal for emotion recognition using audio speech signal features consisting of two functionally independent systems. First, a voice activity detection module (VAD) acts as a filter prior to the emotion classification task. It extracts features from the input audio and uses a SVM classifier to predict the presence of voice activity. Secondly, the speech emotion classifier (EMO) transforms the power spectrum of the signal to a Mel scale and obtains a vector of its characteristics using a convolutional neural network. Emotion labels are assigned using this vector and a KNN classifier. The RAVDESS dataset has been used for training the models obtaining a maximum accuracy of 93.57% classifying 8 emotions.
Index Terms: speech emotion recognition, human-computer interaction, computational paralinguistics

1. Introduction

When people engage in natural conversational interaction, they convey much more than just the meanings of the words spoken. Their speech also conveys their emotional state and aspects of their personality [1]. Paralanguage refers to properties of the speech signal that can be used, either consciously or subconsciously, to modify meaning and convey emotion. Examples of paralinguistic features include those that accompany and overlay the content of an utterance and modify its meaning, such as pitch, speech rate, voice quality, and loudness, as well as other vocal behaviors, such as sighs, gasps, and laughter. Paralinguistic properties of speech are important in human conversation as they can affect how a listener perceives an utterance. Schuller and Batliner [2] presented a detailed survey of computational approaches to paralinguistics, examining the methods, tools, and techniques used to automatically recognize affect, emotion, and personality in human speech. Berkeham and Oguz have very recently presented a detailed survey of emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers for speech emotion recognition [3].

In this paper we propose a speech emotion recognition approach that consists of two independent modules: VAD and EMO. The Voice Activity Detection (VAD) module takes an audio signal as input, analyzes its features by means of a computationally efficient metric evaluation, and classifies its content using a SVM classifier. The result is a binary classification: 1 if it estimates that the audio contains speech activity, 0 otherwise.

Most techniques for speech emotion recognition use Cepstral coefficients in the Mel Frequencies (MFCC), which compute characteristics of both the semantic content of the audio signal (the encoded information) and the contextual content (the state of the interlocutor). The use of Deep Learning techniques for the classification of these characteristics is popular in the field [4, 5, 6]

The speech segments as detected by the VAD module are forwarded to the emotion classifier module (EMO), which transforms the power spectrum computed from the Short-Time Fourier Transform (STFT) of the signal following the Mel scale and obtains a vector of its characteristics using a Convolutional Neural Network (CNN). Transfer learning techniques have been employed to use a pre-trained CNN as a feature extraction model [7, 8]. The resulting feature vector has been used as input to a K-Nearest-Neighbor (KNN) classifier, which ultimately performs the classification task.

The remainder of the paper is structured as follows. Section 2 describes the datasets used to train and evaluate our proposal for speech emotion recognition. Section 3 presents the architecture for the proposed system. Sections 4 and 5 describe the main two modules (VAD and EMO), their practical implementation and the results of their evaluation. Finally, Section 6 presents the conclusions and guidelines for future work.

2. The datasets

Two datasets have been used, one for the VAD module and the other for the EMO module. The dataset used to develop the VAD module was developed by the authors and consists of over 2 hours of recorded speech and non-speech audios on different environments and under different communication contexts. Audio samples were then labeled manually, and the resulting audio files were split into segments of 200ms [9], which has been considered to be a wide enough time window in which perform reliable feature extraction (smaller windows leads to lack of information when computing features, while bigger windows may comprise speech and non-speech subsegments). Additional audio files were generated from the processed dataset using aggregation operations: the samples of every audio segment were shifted in time, compressed and exposed to noise, saving the result as new audios. This operations helps to extend the set of examples without adding excessive redundancy to the dataset. The final dataset contains 150k labelled entries.

For the EMO module the RAVDESS dataset (Ryerson Audio-Visual Database of Emotional Speech and Song) [10] has been used. We have considered 8 emotion categories: calm, happy, sad, angry, fearful, surprise, disgust, and neutral.

This dataset contains labeled files in 3 modality (full AV, video-only and audio-only) and 2 vocal channels (speech and song) from male and female actors. Since our focus was on speech emotion recognition, only audio-only speech samples

10.21437/IberSPEECH.2021-12

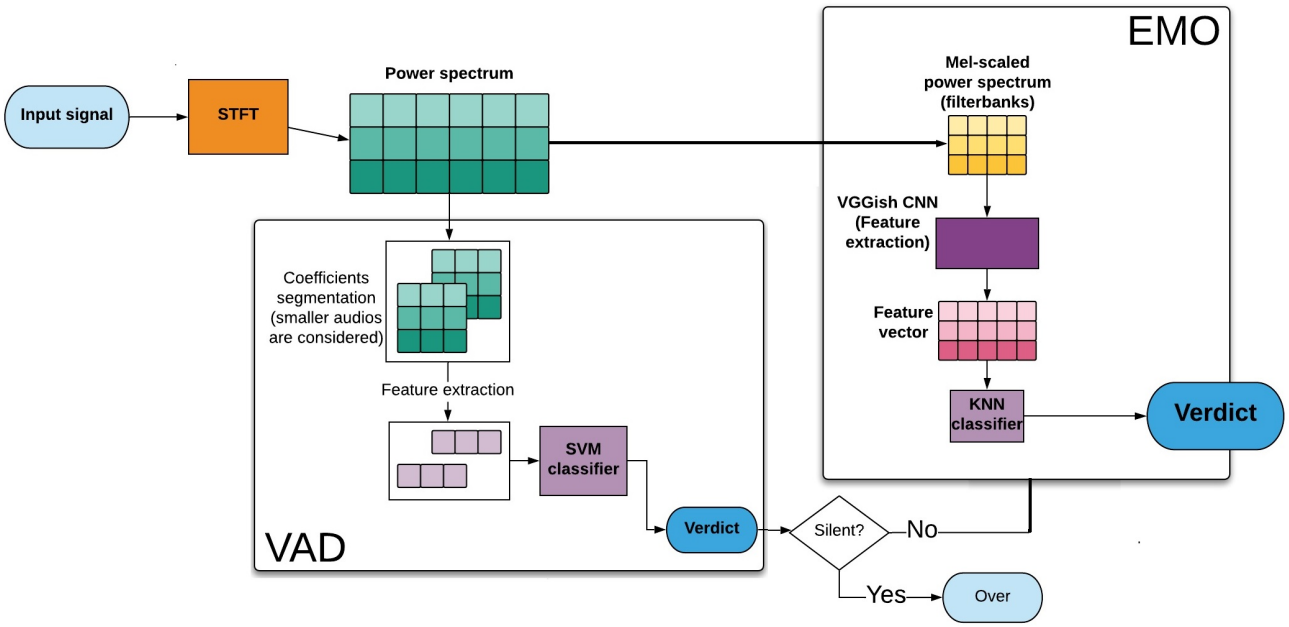


Figure 1: *Proposed architecture for emotion recognition from speech*

have been employed.

The database is gender balanced consisting of 24 professional actors, vocalizing lexicaly-matched statements in a neutral North American accent. Each expression is produced at two levels of emotional intensity (normal and strong intensity), with an additional neutral expression.

The set of 7,356 recordings were each rated 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals who were characteristic of untrained research participants from North America. A further set of 72 participants provided test-retest data.

In order to optimize the training work, the original audios have been segmented into 1s fragments, and each of these segments has been analyzed with the proposed VAD system, eliminating the audios in which there was not speech detected [11, 12, 13]. The audio files in both datasets have a sampling frequency of 16,000 Hz, 32 bits wav-audio format, mono channel.

3. Our proposal

Figure 1 shows the architecture proposed. As it can be observed, the first step consists of obtaining the STFT of the input signal. The STFT operates by splitting the original time signal into multiple smaller segments (frames) and then applying the Fourier Transform on each of them, thus, offering both time and frequency resolution.

The STFT shape and resolution greatly depends on the window features. The implementation used to develop the system allows fine tuning both the window features and shape of the resulting metrics of the STFT analysis, simplifying its usage across the system.

The VAD and EMO modules are described in the following sections. Although they are presented as modules of the proposed architecture, they can also work as functionally independent systems (voice activity detector and speech classifier respectively) that could be plugged into other architectures.

4. The Voice Activity Detection (VAD) module

If the audio segments used do not contain speech information the emotion analysis would be less precise. To avoid this situation, a computationally efficient real time voice activity detection module (VAD) has been implemented and placed as first step before emotion classification. Figure 1 shows the structure of the VAD module: it takes an audio signal as input, frames it into segments of 200 ms, computes the vector of statistical features of each segment and then feeds it to an SVM classifier to generate the final prediction as a binary output (0 if the segment is non-speech or 1 otherwise).

Hit-rate efficiency is achieved by having account of the elements present on each spoken communicative situation: the speaker (different speakers produce audios with different tonal features which are dependent of the individual's anatomy), the speech (the emotional context of the speech can further alter each individual's tonal features) and the environment (noisy conditions lowers the efficiency of the speech detection system).

4.1. Features

The proposed VAD system computes the following statistical metrics:

- **ZCR - Zero-Crossing Rate**: Assuming that the amplitude of a signal is defined in the range [-1, 1], the ZCR coefficient measures the number of times the amplitude changes sign (cross zero) in the time signal. A non-speech signal tends to show a higher ZCR values accounting for the noise being the only element present in the signal: the signal is steadier, and so, oscillates at all times around similar values.
- **HZCRR - High Zero-Crossing Rate Ratio**: While the Zero Crossing Ratio measures the stability of a signal's amplitude over the whole signal, the HZCRR measures the proportion of Zero Crossing Ratio across individual

segments of the signal. Formally is defined as the number of signal segments whose ZCR ratio is above 1.5 fold the average ZCR of the full signal, and represents a localized view of the ZCR.

- **SF - Spectral Flux**: This metric is defined as the mean difference in spectral power between two adjacent frames of the original signal. A higher spectral flux means a higher variance between frames, thus a higher chance for the signal to be speech.
- **STED - Short Time Energy Deviation**: STED measures the variance of the signal level by estimating the difference between the mean and minimum averaged spectral power across the signal frames.
- **SPSD - Short-Term Spectral Power Density**: While the STED coefficient measures the energy difference across time frames, the SPSD follows the same principle applying it to the frequency coefficients of the signal's STFT.
- **BE - Band Entropy**: Entropy measures the amount of information available in a source. The band entropy operates by estimating the entropy across the frequency power coefficients of each frame of the signal's STFT, and then averaging the result by the total number of frames. This feature is computed for multiple frequency bands, thus the original signal must be filtered in order to obtain representations of different frequency regions.
- **BP - Band Periodicity**: This metric is computed measuring the correlation of the waveform of adjacent signal frames. Like band entropy, this feature highly benefits from the multiband signal representation (human voice tends to oscillate in specific frequency bands), thus a filtered signal must be provided to properly compute the metric.

4.2. Classification

The VAD system operates on smaller segments of the original signal, thus the previously described features are computed multiple times for every audio signal that is feed into the system. Once all the feature coefficients are resolved, the VAD system uses the resulting feature vector as the input of a classifier, which ultimately performs the audio activity estimation.

For the classification task two base models have been considered: Random Forest (RF) and Support Vector Machines (SVM) [14, 15].

In order to adjust the models correctly, different critical settings have been considered for each of the classifiers and compared using a cross-validation study with the samples generated. The optimal configurations obtained for the classifiers were: Random Forest (200 trees, quadratic selection of characteristics), linear SVM (C=1000) and polynomial SVM (C=2000).

Figure 2 shows the accuracy results obtained with the selected SVM classifier.

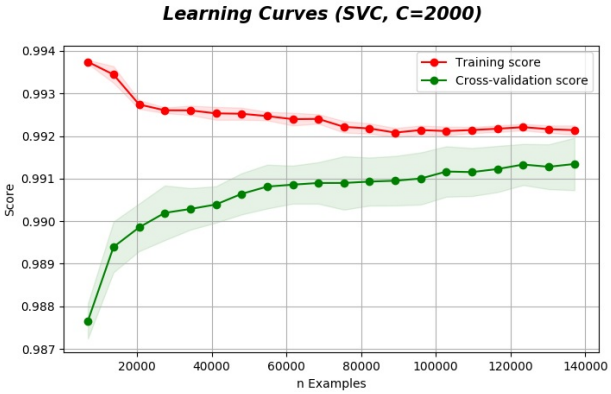


Figure 2: *Results obtained for the VAD module with the selected SVM classifier*

5. The emotion classifier module (EMO)

As Figure 1 shows, the first step in the proposed emotion classifier is to transform the power spectrum computed from the STFT of the signal following the Mel scale.

Then, MFCC coefficients are taken as input into the VGGish model to extract the input features used for the emotion recognition[1]. VGGish is a CNN model developed by the Sound Understanding team of Google to compute 128-dimension features vectors from Mel scaled input spectrograms. The model mixes multiple convolution-pooling layers to produce the feature vector.

The model consists of several convolution layers with pooling on the maximum value in the output of each one of them, consisting of a last layer totally connected for classification.

5.1. Transfer Learning

Implementing a CNN architecture specifically built to recognize emotions from a given input vector requires high amounts of data to generate an efficient model. Transfer learning techniques can greatly optimize the generation of Deep Learning models by using the network structure implemented in an existing model (with similar classification functions to the desired model), removing its output layer and implementing one specific to the classification task in hand. The resulting model can then be trained with the relevant data to fit the wanted model. Alternatively, instead of a classification network, the reference model can be used to extract features from an input vector, and then use the resulting feature vector as input for another Machine Learning classifier, which ultimately performs the classification task.

To use the VGGish model in the proposed application, a standard CNN model has been built with the same layer architecture used by VGGish. The last layer has been replaced by a Machine Learning classifier, which performs the emotional classification task on the characteristics extracted by the neuron network.

The training of the models was completed using the RAVDESS dataset, segmenting the original audios into fragments of one second and eliminating the audios in which there was not a sufficient amount of speech. Additionally, aggregation operations have been applied to the resulting audios (audio displacement, compression and noise introduction) to extend the number of available examples, which has made it possible to increase to a total of 30,000 files preserving the nature of the emotion that they represent.

[1]The definitions of the VGGish model implemented in Keras with tensorflow backend are available at https://github.com/DTaoo/VGGish. Last access: February 2021

Table 1: *Results of the evaluation of the set of classifiers for the EMO module*

Alternative	Model	Hyperparameters	Precision	Delta
CNN	KNN	N = 3	93.57%	3.2%
CNN	SVC	Penalty = 1.75	93.69%	5.67%
CNN	MLP (1) Layers: (80);	Learning rate: 0.05	92.66%	7.34%
CNN	MLP (2) Layers: (80, 80);	Learning rate: 0.05	93.24%	6.76%
MFCC	SVC	Penalty = 1.0	93.11%	5.3%
MFCC	KNN	N = 2	92.36%	7.64%

5.2. Classification

Starting from the log-Mel scale power spectrum, the resulting coefficients are supplied as input to the proposed VGGish model. The model extracts the characteristics of the input data, being these coefficients those that are finally supplied to the classifier for the detection of emotions. For the classification task of the VGGish output, 3 different models have been considered:

- KNN - K Neighbors: The model uses the concept of majority voting fixing classes to each classifiable object depending on the class of its neighbors: the most common neighbor class is the class assigned to the element.
- SV - Support Vector Machines: Support Vector Machines recursively transforms the dimensionality of the inputs to find hyperplanes that splits and classifies the whole space.
- MLP - Multilayer Perceptrons: Represents a neural structure formed by interconnected layers of nodes. The input is transformed while propagating through the layers to the output, where is finally classified.

As with the VAD module, multiple configurations have been considered on each classifier to select the one providing the best results. Table 1 shows the results obtained with the different classifiers. The Delta parameter refers to the differences between the test and validation sets.

All the selected models provide satisfactory results, however, the overfitting shown by the MLP models of the CNN alternative, as well as the KNN classifier considered during the classification of the MFCC coefficients, discards them as candidates for the final emotion classifier. Therefore, the model chosen for its integration in the system is the KNN, which shows a fairly close convergence between the test and training sets as well as offering an accuracy close to 94%. This result improves the baselines obtained for the RAVDESS corpus in recent proposals [16, 4, 14]. Figure 3 shows the confusion matrix obtained for this classifier.

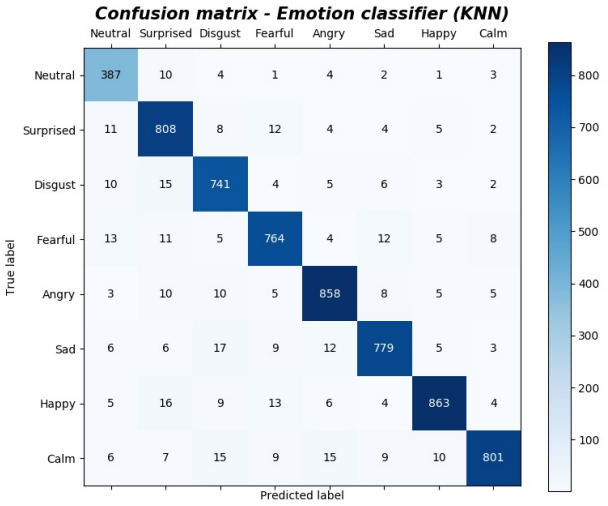


Figure 3: *Confusion matrix for the KNN classifier*

6. Conclusions and future work

In this paper we have presented a proposal for emotion recognition using speech features, transfer learning and convolutional neural networks. The proposed architecture integrates two independent subsystems that collaborate in the task of classifying an audio according to its content: VAD and EMO.

The VAD subsystem determines the existence of speech components in the audio using a support vector machine, which takes the metrics previously computed by the subsystem and decides on the presence of voice in the signal.

The EMO subsystem (classifier of emotions), acts on the basis of the decision reached by the VAD and transforms the power spectrum of the signal to a Mel scale to obtain a vector of its characteristics using a network of convolutional neurons. Then, an emotion tag is assigned by a KNN classifier.

We have implemented our proposal and used it with the RAVDESS dataset, obtaining an accuracy of 93.57% with 8 emotion categories. As future work we plan to evaluate our proposal with more datasets.

7. Acknowledgements

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823907 (MENHIR project:https://menhir-project.eu) and by the Spanish Ministry of Economy, Industry and Competitiveness through CAVIAR (MINECO, TEC2017-84593-C2-1-R).

8. References

[1] M. McTear, Z. Callejas, and D. Griol, *The Conversational Interface: Talking to Smart Devices*. Springer, 2016.

[2] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. Wiley, 2013.

[3] M. B. Akçay and K. Oguz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56 – 76, 2020.

[4] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.

[5] L. Sun, B. Zou, S. Fu, J. Chen, and F. Wang, "Speech emotion

recognition based on DNN-decision tree SVM model," *Speech Communication*, vol. 115, pp. 29–37, 2019.

[6] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[7] Z. Ahmad, R. Jindal, A. Ekbal, and P. Bhattachharyya, "Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding," *Expert Systems with Applications*, vol. 139, p. 112851, 2020.

[8] J. C. Hung, K.-C. Lin, and N.-X. Lai, "Recognizing learning emotion based on convolutional neural networks and transfer learning," *Applied Soft Computing*, vol. 84, p. 105724, 2019.

[9] B. Atmaja and M. Akagi, "Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model," in *Proc. of IEEE International Conference on Signals and Systems (ICSigSys)*, Bandung, Indonesia, 2019, pp. 40–44.

[10] S. Livingstone and F. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," *Plos One*, vol. 13, no. 5, p. e0196391, 2018.

[11] F. Saki and N. Kehtarnavaz, "Emotion Detection Using MFCC and Cepstrum Features," in *Proc. of the 4th International Conference on Eco-friendly Computing and Communication Systems*, Orlando, FL, USA, 2015, pp. 29–35.

[12] S. Lalitha, D. Geyasruti, R. Narayanan, and M. Shravani, "Automatic switching between noise classification and speech enhancement for hearing aid devices," in *Proc. of 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, Florida USA, 2016, pp. 29–35.

[13] Z. Tuske, P. Mihajlik, Z. Tobler, and T. Fegyo, "Robust voice activity detection based on the entropy of noise-suppressed spectrum," in *Proc. of 9th European Conference on Speech Communication and Technology (Interspeech)*, Lisbon, Portugal, 2005, pp. 245–248.

[14] A. Bhavan, P. Chauhan, Hitkul, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Systems*, vol. 184, p. 104886, 2019.

[15] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction," *Information Sciences*, vol. 509, pp. 150–163, 2020.

[16] M. Kwon and S. Kwon, "A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition," *Sensors*, vol. 20, pp. 1–15, 2019.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Using Audio Events to Extend a Multi-modal Public Speaking Database with Reinterpreted Emotional Annotations

Esther Rituerto-González, Clara Luis-Mingueza, Carmen Peláez-Moreno

Signal Theory and Communications Department University Carlos III of Madrid, Spain

erituert@ing.uc3m.es, cluis@pa.uc3m.es, carmen@tsc.uc3m.es

Abstract

Emotions present in speech provide a lot of information about the emotional state of a speaker. Affective Computing is an emerging field that analyses these states and tries to improve human-computer interaction tasks.

In this paper we aim to present a preliminary study on the analysis of stress in speech and acoustic events that may possibly cause it. We merge four speech & audio technologies: speaker and emotion recognition and acoustic event detection and classification, and explore how they influence each other.

We perform initial experiments on BioSpeech, a multi-modal emotions database we have extended with acoustic events and discuss a novel labelling process targeted to improve the classification performance.

The current study is intended as a classification and detection baseline for the mono-modal speech tasks described, and presents a discussion on the future work and multi-modal architectures to be implemented in a cyberphysical system for gender-based violence automatic detection.

Index Terms: affective computing; speaker recognition; acoustic event detection; acoustic scene analysis

1. Introduction

Acoustic Scene Analysis and Interpretation aims to process and interpret the acoustic information in the environment usually captured by a multi-microphone acquisition system. In this paper, we are concerned with a specific type of acoustic scene: that in which Gender-based Violence (GV) appears. Indeed, GV is one of the biggest social problems in the world. Its eradication is essential for achieving gender equality, the fifth of the Sustainable Development Goals (SDG) adopted by all UN Member States in 2015, as part of the 2030 Agenda for Sustainable Development.

In particular, we aim at *detecting* this kind of scene by using Bindi [1, 2], a multimodal cyberphysical system. Bindi uses bio-signal processing technologies, wearable edge computing, and machine learning –among other disciplines. Unlike other technological solutions that mainly focus on providing the means for the victims to call for help (camouflaged panic buttons with geolocalization [3], voice-activated alert devices, etc.), Bindi employs intelligent bio-signal processing for affective computing to *autonomously detect a violent situation*. These bio-signals are collected by smart sensors, including a microphone, integrated in wearable edge devices.

In this preliminary paper, we will limit our scope to the processing of the auditory modality, leaving its effective integration with the physiological signals ([4, 5]) for further developments. It is important to remark that the complexity and energy restrictions typical of cyberphysical systems prevent the uninterrupted collection of this auditory information. In our setup, the microphone is woken up by an alarm from the physiological signals[1]. This alarm is configured in a conservative way with a high false alarm rate to prevent misses. Therefore, the auditory information is used to provide the *context* necessary to *disambiguate* the information carried by the physiological signals.

Thus the challenges we face in this paper are: first, the lack of appropriate datasets for this task that combines physiological and auditory signals and second, the interaction between speech and audio technologies in our setup. In particular, we describe a principled means to extend an existing multi-modal emotions database with violence related audio events and initial experiments to assess their effects on speaker and stress detection tasks on the one hand, and acoustic event detection and classification, on the other.

This paper is organized as follows: first, we outline the state of the art in Sec. 2, our reinterpretation and extension of the targeted dataset is in Sec. 3 followed by experiments and results in Sec. 4, and closing discussion and further work in Sec. 5.

2. Related work

Few studies investigate the relationship between acoustic events and the elicitation of fear [6], and there are no databases –to the knowledge of the authors– that include acoustic events and speech recorded consequentially. A large-scale dataset of manually annotated audio events is AudioSet [7].

As for databases containing real-life speech under panic or fear circumstances, literature is scarce, but at the moment there are a few in which stressed speech is either simulated or recorded under real conditions, such as SUSAS [8], UT-Scope [9], VOCE [10] and BioSpeech [11]. As films are one of the most effective ways to elicit emotions [12], our team UC3M4Safety is currently recording a database for fear recognition in the context of gender-based violence in the EMPATIA project [13].

Stress is barely considered an emotion although it is intimately related to anxiety and nervousness, being closely connected to fear. Among the measurable physiological consequences of stress appear respiratory changes, increased heart rate, skin perspiration, and increased muscle tension of the vocal folds and vocal tract [14].

The most suitable database to our interests is BioSpeech (BioS-DB) [11], since it includes continuous-time annotations in the Arousal/Valence space for non-acted speech presumably stressed due to its public speaking setting, and incorporates physiological data (Blood Volume Pulse –BVP– and Skin Conductance –SC–) as in Bindi, which could be of great use for multi-modal models in the future. Note, however, that the purpose of this dataset collection is quite different from ours since their creators aimed at predicting bio-signals from speech.

[1] There are also short wake up periods to check the functionality and update certain parameters periodically.

10.21437/IberSPEECH.2021-13

In general, the main difficulty with emotionally labelled data relies on the proper labelling process. There is no universal agreement on how to categorize or measure emotions. The self-assessment measures annotations by a specific subject can differ from labels annotated by external evaluators observing the said subject. We will further discuss the impact of this issue on our reinterpretation of BioS-DB in Section 3.1.

On the other hand, the emotional state of the subject could influence negatively the performance of any speech technology and in particular, their identification. This constitutes a challenge we will refer as to *affective speaker recognition*. Tracking the user's voice separating it from the rest of the speakers opens an interesting possibility for situations where it would be desirable to identify all the speakers involved in the scene, e.g., in case of legal evidence required.

In this study we present a synthetic augmentation of BioS-DB with acoustic events that most likely could cause an stressful reaction loosely synchronized with the time instants where the labels denote an acute stress occurrence. We introduce a reinterpretation of the labelling of BioS-DB, more suitable for our classification task. Moreover, we introduce the problem of the relation of acoustic events and emotions, and use shallow deep learning models to establish a baseline for mono-modal emotion (ER) and speaker recognition (SR), and a pretrained deep learning model for acoustic events detection (AED) and classification (AEC) tasks.

3. Methodology

3.1. Relabelling of BioSpeech

BioS-DB [11] is a multi-modal public speaking database which includes continuous-time emotional annotations. It consists of 55 speakers reading two texts, one in German and one in English, while their physiological variables (BVP, SC) and speech are being recorded. This database responds to the idea that performance anxiety can happen when speaking aloud and can be reflected in the physiological variables and speech. Three annotators with previous training use a joystick to obtain continuous time labels for the emotional state of the speaker in a 2D space of which their axis represent *arousal* and *valence*.

The authors of BioS-DB used the evaluator weighted estimation (EWE) for computing the collective time-continuous ratings when creating a gold standard for the emotional labels from the three individual time-continuous annotations [15].

Though EWE is reliable when the number of annotators is large, in this case the possibility of disparity in the ratings is very high. The subjective evaluation of each scorer affects their ratings, besides the bias of the possible comparisons between consecutive speakers. These factors can induce to a lot of variability and discrepancies, and a weighted combination of the labels of each annotator may not be the optimal merging method. This was specially damaging for our purposes, that are different from those of the creators of the dataset.

Thus, we propose a re-labelling of BioS-DB Arousal and Valence values quantizing them into 4 categorical quadrants. This is crucial to define a classification task instead of using a regressor. These four quadrants are:

- High Arousal, High Valence: Excitement (Q1)
- High Arousal, Low Valence: Stress (Q2)
- Low Arousal, Low Valence: Sadness (Q3)
- Low Arousal, High Valence: Calmness (Q4)

We also believe that although BioS-DB counts with a very precise temporal resolution in the labelling, coarser time resolution for capturing the underlying emotions in speech is more suitable in classification tasks such as ours.

In particular, the raw annotations in BioS-DB from each annotator were originally sampled at 2Hz and their range was [-1000, 1000]. Therefore for our purposes, we downsample the signals to 1Hz to obtain one label per second, which will be our baseline working frequency for future data fusion schemes. To compute a combined final label for each second, we chose the two annotators that had labelled closer in the 2D space[2], and based on the sign of the Arousal and Valence values, we convert these into a categorical label in each of the four quadrants. If the quadrant where the two labels considered lay coincides, it is chosen as the aggregated label value, otherwise, we assign an undetermined value, x.

Then, we analyze several cases for the undetermined labels: if x is due to a transition between quadrants (one annotator has crossed the boundary but the other has not yet), we randomly choose any of the two quadrants. Otherwise, we consider whether two annotators fall into the same quadrant even though they are not the closest in the 2D space. If so, the aggregated label is the corresponding to that quadrant. This process solves a great amount of undetermined labels. For the rest and those cases where we found several x in a row, we used a 5-second window and replaced the unknown labels with majority voting.

Our process takes into account the proximity of the labels of the raters, which provides confidence about the resulting label since the annotators interpret the 2D space in terms of the quadrants meaning. Transitions between quadrants are considered carefully since people do not leap from one emotional state to another suddenly. The smoothing window provides a smooth label signal by avoiding sharp changes between quadrants.

Finally, for our task of automatic detection of gender-based violence situations, the second quadrant $Q2$ where emotions related to stress, anxiety and fear rely, will be chosen as target. Thus, for the baseline experimentation we considered two types of labellings: quadrants and binary (considering $Q1$, $Q3$, $Q4$ as the negative label, and $Q2$ as the positive).

	Q1	Q2	Q3	Q4
Original BioS-DB	29.22	22.56	8.53	39.67
Reinterpreted BioS-DB	22.16	39.04	8.56	30.24

Table 1: *Percentage of labels in each quadrant*

3.2. BioSpeech+

As stated in previous sections, the ultimate goal of Bindi is to provide an autonomous and inconspicuous tool to detect Gender-based Violence. Regarding speech and audio, we aim at tracking and identifying the user's voice [1] and then use it to detect fear or panic. To improve the precision of the system, this is contextualized by the analysis of the acoustic scene (background sounds and noises) by using a Sound Event Detection and Classification (AED/C) system.

BioS-DB is being used as a proxy to our problem. However, for our specific purposes it is key to complement the spoken information with knowledge about the events present in the acoustic scene: in many cases, panic could cause a GV victim to remain in silence. That is why environmental sounds, that is, the characterization of the acoustic scene, may provide useful information for the detection system.

Therefore we introduce a preliminary procedure to extend the BioSpeech database, consisting of the original speech audio

[2]Preliminary analysis considered selecting labels in terms of 1) proximity or 2) quadrant concordance but experimental results proved that the first approach was more consistent and stable in time.

```
for each lang {'de' or 'en'} do
    for file in lang_foreground_path do
        compute file duration;
        define Scaper object {sample rate = 16 kHz, n_channels =
          1, set ref_db (loudness level)};
        reset previous event specifications;
        groupby: sequential Q2 labels (binary) from correspondent
          BioS-DB.csv file;
        for each Q2 group do
            define event_duration and start_time from Q2 labels;
            if binary_label == 1 then
                add background event fixing {event_duration,
                  start_time};
            end
        end
        add foreground event fixing {file};
        synthesize defined mix;
    end
end
```

Algorithm 1: Procedure for mixing BioSpeech and Audioset samples with Scaper

files synthetically enriched with environmental sounds. The process is an initial approach open for discussion.

We make use of AudioSet [7], a large-scale collection of human-labeled 10-second sound clips drawn from YouTube videos. Audioset provides 2,084,320 samples containing 527 weak annotations at clip level of sound events. We have selected a subset of 2108 samples from Audioset, belonging to 83 classes, to extend the original BioS-DB. To choose classes related to violent events, we have employed the audiovisual stimuli selected in the early stages of the UC3M4Safety dataset collection (in progress). The initial selection was made by experts in VG and later on validated by more than 1300 volunteers [13]. To identify the acoustic events present in the audiovisual stimuli we have employed a pre-trained sound event classification model: YAMNet [16].

YAMNet is a Convolutional Neural Network (CNN) pre-trained on 521 classes of AudioSet, ready to perform inference over audio files to classify occurring sound-events. At the preprocessing stage, the audio signal is normalized, and converted to a 16 kHz mono. Then a log-mel spectrogram of 64 bins is computed to extract a time-frequency representation of the audio signal as an image. These features are fed in patches of 0.96s to the network. The inference stage provides the final score averaged over all the input frames, time-dependent output scores of each class for every 960 ms of raw audio data. It also allows extracting 1024-dimensional embeddings corresponding to the activations of the top convolutional layer.

Regarding the synthetic mixing, the process is based on the data-augmentation pipeline followed in Task 4 of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 challenge [17]. Scaper [18] allows us to define probability distributions for the occurrence and duration of the sound events. Thus, the system generates as many synthetic mixes as desired from audio previously classified as foreground or background. In our particular case, foreground events are the original BioS-DB samples and background events are the samples of the Audioset subset. The number of generated mixes has been set to 110: we generate one mix per BioS-DB file, considering recordings captured by the lavalier microphone, i.e. 55 German and 55 English-speaking *audio1* recordings.

The details about the mixing procedure, taking into account the new binarized labels explained in Section 3.1, is presented in pseudocode format in Algorithm 1.

The rationale for this is to provide a non-deterministic relationship between stressing sounds and the appearance of stress in the speaker. In addition to managing probability

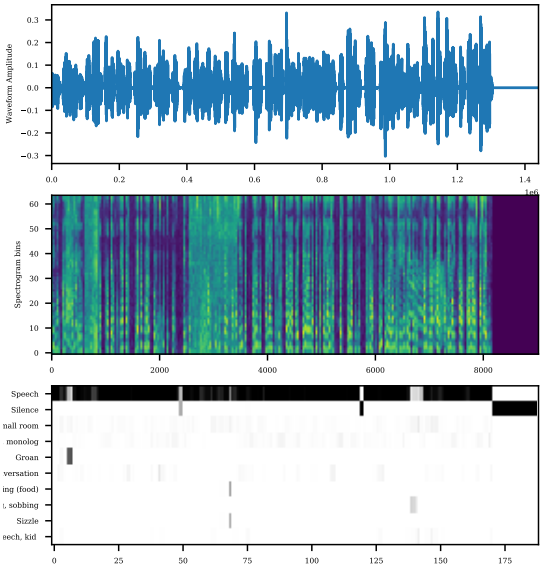


Figure 1: *YAMNet output of a sample of BioSpeech+ with the temporal representation (top), spectrogram (middle, bands spanning 125 to 7500 Hz) and top events found (bottom)*

distributions and timing of the events, Scaper allows pitch shifting and time stretching operations over foreground samples, that could be used for further augmenting the dataset.

4. Preliminary Experiments and Results

To find out if we could use AED/C of the background events to assist the Speaker (SR) and Emotion Recognition (ER) tasks, we have combined them at different SNR[3] (-5, 5 and 15 dB).

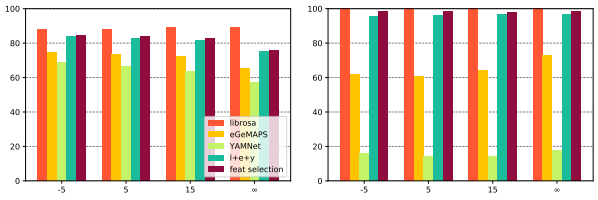
We extracted features from well-known libraries used for SR, ER and AED/C, respectively: librosa [19], eGeMAPS [20] from the openSMILE toolkit [21] and YAMNet embeddings [16]. The size of our working window is one second. This is a compromise between computational complexity and speed and a requirement in Bindi. Thus, from librosa we extracted 19 features with a window size of 20ms and a 10ms overlap and then their mean and standard deviations every second resulting in 38 features per second. Using openSMILE we extracted the eGeMAPS feature set with 88 features. For extracting features suitable for audio events we used the 1024-dimensional embeddings corresponding to the activations of the top convolutional layer of YAMNet. A feature selection method where the correlation of the concatenation of the three feature sets was used to remove the features with a correlation higher than 95%. This resulted in a reduction of the 68% of the features. Examining the correlation matrices we confirmed that most YAMNet features were highly correlated with each other. All features were standardized by using z-score normalization.

With the chosen window size, BioS-DB contains approximately 5000 samples. This is a small size for the use of deep neural networks, so a simple Multi-Layer Perceptron (MLP) implemented with scikit-learn [22] and two shallow architectures implemented with Keras [23] were tested, working towards maintaining a low computational complexity. The first of them consists of two hidden fully-connected layers. The second is a combination of a convolutional 1D layer, a bidirectional GRU layer and a fully-connected layer. This model responds to the idea that it is important to extract

[3]For the SNR measure we consider the foreground speech from BioS-DB as the 'signal´ and the violent audio events as 'noise´.

Model	LIBROSA	*p*	EGEMAPS	*p*	YAMNET	*p*	L+E+Y	*p*	FEAT SEL	*p*
EMOTIONS RECOGNITION BINARY										
MLP	89.1±0.9	*12k*	65.4±1.8	*27k*	57.2±1.4	*307k*	75.3±1.7	*345k*	75.8±1.3	*111k*
K2D	82.4±1.0	*3k*	54.2±0.8	*5k*	32.7±9.0	*52k*	66.3±1.4	*58k*	65.1±1.2	*19k*
KCGD	80.9±1.8	*9k*	54.3±2.7	*12k*	30.4±5.6	*72k*	66.7±1.3	*80k*	67.2±1.3	*30k*
EMOTIONS RECOGNITION 4-Q										
MLP	90.0±0.9	*12k*	45.5±1.1	*27k*	35.8±1.7	*307k*	59.5±1.0	*346k*	60.4±1.6	*112k*
K2D	73.2±1.0	*3k*	47.7±2.0	*6k*	37.6±1.0	*52k*	56.8±1.0	*59k*	57.8±1.2	*19k*
KCGD	73.2±0.9	*9k*	47.9±1.0	*12k*	37.6±0.9	*72k*	58.7±1.2	*80k*	56.9±1.7	*30k*
SPEAKER RECOGNITION										
MLP	100±0	*28k*	72.7±0.6	*43k*	17.8±1.4	*324k*	96.4±1.0	*361k*	98.35±0.3	*128k*
K2D	99.9±0.1	*4k*	64.3±2.0	*7k*	15.21±1.4	*53k*	95.9±0.8	*60k*	96.6±0.7	*20k*
KCGD	100±0	*10k*	50.9±0.7	*13k*	12.6±1.9	*73k*	90.8±1.3	*81k*	95.7±0.9	*31k*

Table 2: *F1-score results for clean BioS-DB. MLP refers to the Multi-Layer Perceptron, K2D refers to the 2-dense layers model in Keras and KCGD refers to the Keras model composed of a Convolutional 1D, Bidirectional GRU and Dense layers. Mean and standard deviation results are shown for a 5-fold validation.*



(a) Emotion Recognition *(b) Speaker Recognition*

Figure 2: *F1-score results with Multi-Layer Perceptron*

information from the temporal context distribution of the features data. The models were compiled using Adam with a learning rate of 0.001, categorical cross-entropy as the loss function and f1-score as the metric to evaluate performance due to the imbalance of the dataset. For all experiments we used a 5-fold cross-validation.

The results for BioS-DB without the audio events are shown in Table 2, where p represents the number of parameters of each model. For the three tasks under consideration, MLP with librosa achieves the best performance. It is worth noting that librosa features achieve the maximum score for the SR task.

The differences in performance between features can be due to multiple reasons: their nature –librosa and eGeMAPS features are manually extracted whereas YAMNet's are automatically extracted from a pretrained sound-event detection network–, their number –38, 88 and 1024 respectively–, and their specific potential to represent emotions or speaker information. Further examination is certainly needed.

Figs. 2 (a) and (b) provide the results for ER and SR respectively for different SNRs. Specifically, Fig. 2 (a) shows the results for binary ER for the model that performed the best (MLP). All the feature sets –except maybe librosa, which remains stable– show a trend to improve the f1-score as the SNR value gets lower, that is, when the acoustic events overlay the speech. This demonstrates that extending our database with stressful events comes in handy for the recognition of stress in speech. All the feature sets, in a greater or lesser extent, are able to capture information about the acoustic events which are considered stress triggers.

As for Fig. 2 (b) we can observe an almost perfect performance for librosa features, and a considerable decrease in efficiency for YAMNet embeddings. However, the performance decreases with lower SNRs contrary with what we observed in Figure 2 (a). This means that acoustic events do not facilitate the SR task. Besides, YAMNet embeddings do not seem to capture relevant information about the acoustic cues of the speech that could help distinguish between speakers.

A for the AED/C task, we illustrate an example in Fig. 1 of the performance of YAMNet classifying a 90 s mixed audio. The dataset has been pre-processed to match YAMNet's requirements ($f_s = 16KHz$, mono, amplitude normalized to $[-1, 1]$) and then fed into the model. The only free parameter is `patch_hop`, which was set to $0.48s$. The audio corresponds to a women reading a text in German. When the first Q2 annotations occur, a background event of a man groaning can be heard. Some yelling and a snoring occur right after but the network only captures it by decreasing the confidence on the speech class. The next background event is a wind noise that is misslabeled as 'Frying (food)'. Lastly, a man whining sound is classified as crying/sobbing with low confidence.

5. Dicussion and Future Work

We draw from the premise that detecting violent situations involves taking into account speech and acoustic contexts since they could be correlated. However, there are no non-acted datasets that allow to elicit this relationship. In Section 3.1 we reinterpreted BioS-DB labels. The samples labelled $Q2$ were interpreted as those related to fear, anxiety or stress, but we should note that without the *dominance* dimension, emotions such as anger or rage could lay in that quadrant too. Using those labels we have extended BioS-DB with stressful sound events, as described in Section 3.2.

In this preliminary study we focused on speaker, stress, and acoustic events in the background. Both the feature sets and algorithms were used with the aim of keeping low the computing load and taking into account the number of samples of the database used. Stressful acoustic events with a non-deterministic correlation to stressed speech utterances proved to be beneficial to some extent for the classifications of binary emotional utterances. On the contrary, they were not helpful (eGeMAPS or YAMNet) or irrelevant (librosa) in the recognition of the speaker.

This research leaves many open questions and future lines of work. Since Scaper allows us to define probability distributions for the appearance and duration of the sound events, the procedure defined in Section 3.2, ready to perform the addition of background events when `binary_label` is $Q2$, could be extended by proceeding in a similar way with non-stressful events whenever `binary_label` is not $Q2$, making the resulting mix sound more realistic.

Also background sounds in the mixing process can be adapted to any kind of problem, resulting into new combinations of the BioS-DB and other datasets. As the main goal of Bindi is to detect and prevent Gender-based Violence, these background events could correspond to audio clips of movie scenes representing a GV scenario, selected with expert knowledge and guidance. This way it could be possible to count with a synthetic dataset of Gender-based Violence situations or other different kinds of situations.

6. Acknowledgements

This work has been partially supported by the Dept. of Research and Innovation of Madrid Regional Authority, in the EMPATIA-CM research project (reference Y2018/TCS-5046). We thank NVIDIA for the donation of the TITAN Xp. The authors also thank the Spanish Ministry of Science, Innovation and Universities (MCIU) for the FPU grant FPU19/00448 and the rest of the members of UC3M4Safety for their support.

7. References

[1] E. Rituerto-González, A. Gallardo-Antolín, and C. Peláez-Moreno, "Speaker Recognition under Stress Conditions," in *Proc. IberSPEECH 2018*, 2018, pp. 15–19. [Online]. Available: http://dx.doi.org/10.21437/IberSPEECH.2018-4

[2] E. Rituerto-González, A. Mínguez Sánchez, A. Gallardo-Antolín, and C. Peláez-Moreno, "Data augmentation for speaker identification under stress conditions to combat gender-based violence," *Applied Sciences*, vol. 9, p. 2298, 06 2019.

[3] M. de Igualdad. (2020) Telematic control devices of measures and withdrawal penalties. [Online]. Available: https://violenciagenero.igualdad.gob.es/informacionUtil/recursos/disposi-tivosControlTelematico

[4] J. A. Miranda-Calero, R. Marino, J. M. Lanza-Gutierrez, T. Riesgo, M. Garcia-Valderas, and C. Lopez-Ongil, "Embedded emotion recognition within cyber-physical systems using physiological signals," in *Conf. on Design of Circuits and Integrated sys. (DCIS)*, 2018.

[5] E. Rituerto-González, J. A. Miranda, M. F. Canabal, J. M. Lanza-Gutiérrez, C. Peláez-Moreno, and C. López-Ongil, "A hybrid data fusion architecture for bindi: A wearable solution to combat gender-based violence," in *Multimedia Coms., Services and Security*. Springer Intl. Publishing, 2020, pp. 223–237.

[6] T. Garner and M. Grimshaw, "A climate of fear: Considerations for designing a virtual acoustic ecology of fear," in *Proceedings of the 6th Audio Mostly Conference: A Conference on Interaction with Sound*, ser. AM '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 31–38. [Online]. Available: https://doi.org/10.1145/2095667.2095672

[7] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, *"Audio set: An ontology and human-labeled dataset for audio events"*.

[8] J. Hansen and S. Bou-Ghazale, "Getting started with susas: A speech under simulated and actual stress database," vol. 4, 01 1997.

[9] A. Ikeno, V. Varadarajan, S. Patil, and J. H. L. Hansen, "Ut-scope: Speech under lombard effect and cognitive stress," in *2007 IEEE Aerospace Conference*, 2007, pp. 1–7.

[10] A. Aguiar, M. Kaiseler, M. Cunha, J. Silva, M. H., and P. Almeida, "Voce corpus: Ecologically collected speech annotated with physiological and psychological stress assessments." 05 2014.

[11] A. Baird, S. Amiriparian, M. Berschneider, M. Schmitt, and B. Schuller, "Predicting biological signals from speech: Introducing a novel multimodal dataset and results," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 2019, pp. 1–5.

[12] Y. Deng, M. Yang, and R. Zhou, "A new standardized emotional film database for asian culture," *Frontiers in Psychology*, vol. 8, p. 1941, 2017. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2017.01941

[13] M. Blanco-Ruiz, C. Sainz-De-Baranda, L. Gutiérrez-Martín, E. Romero-Perales, and C. López-Ongil, "Emotion elicitation under audiovisual stimuli reception: Should artificial intelligence consider the gender perspective?" *International Journal of Environmental Research and Public Health*, vol. 17, pp. 1–22, 11 2020.

[14] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Transactions on Affective Computing*, vol. PP, pp. 1–1, 07 2019.

[15] S. Hantke, E. Marchi, and B. Schuller, "Introducing the weighted trustability evaluator for crowdsourcing exemplified by speaker likability classification," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 2156–2161. [Online]. Available: https://www.aclweb.org/anthology/L16-1342

[16] M. Plakal and D. Ellis, "Yamnet," https://github.com/tensorflow/models/tree/master/research/audioset/yamnet, Accessed: 2020-12-30.

[17] N. Turpault, R. Serizel, P. Shah, J. Salamon, and A. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis." [Online]. Available: https://hal.inria.fr/hal-02160855v2

[18] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2017, pp. 344–348.

[19] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," 01 2015, pp. 18–24.

[20] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, pp. 1–1, 01 2015.

[21] F. Eyben, M. Wöllmer, and B. Schuller, "opensmile – the munich versatile and fast open-source audio feature extractor," 01 2010, pp. 1459–1462.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015, Accessed: 2020-12-30.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Query-by-Example Spoken Term Detection using Attentive Pooling Networks at ALBAYZIN 2020 Evaluation: The AUDIAS-UAM System

Juan Ignacio Álvarez-Trejos[1], *Doroteo T. Toledano*[1]

[1]AUDIAS - Audio, Data Intelligence and Speech
Universidad Autónoma de Madrid

juani.alvarez@estudiante.uam.es, doroteo.torre@uam.es

Abstract

Query-by-example Spoken Term Detection (QbE-STD) is a key technology to harness the large amount of audiovisual content that is being stored and generated nowadays. Using audio example queries for STD has several advantages such as requiring less resources (both computational and linguistic) and resulting in less language-dependent systems. A further advantage is the possibility of developing neural end-to-end models. In this paper, we explore one of these models for QbE-STD. The model starts projecting the input pair formed by a query and a segment into fixed-length vector representations. Then, a distance between these vectors is calculated to generate a detection score. To learn similarities over the projected input pair, a two-way attention model, called attentive pooling networks, has been used. Both elements in the input pair can influence the vector representation of the other, paying more attention to the frames that contain key information of both the query and the occurrence. Our main objective is to explore if this model can find similarities regardless of the language used for training. We start showing the effectiveness of the proposed model on the Librispeech corpus, and then we evaluate it on the ALBAYZIN 2020 Search-on-Speech evaluation data.

Keywords: Query-by-example, Spoken term detection, End-to-end systems, Two-way attention, Attentive pooling networks

1. Introduction

Spoken term detection (STD) is defined as the task of retrieving audio segments which contain a query from an audio archive. If the query is in text form, i.e. keyword search [1], the task is typically solved based on automatic speech recognition (ASR) technology. If the query is an acoustic example (spoken query) we have Query-by-Example (QbE) STD. The QbE-STD task can be formulated as a detection task in which the input is an audio pair and the output is the list of hypothesis detected, which contains a detection score and the time intervals in which the detection resides. The most attractive feature of QbE-STD is that it is not necessary to transcribe the audios to text, and with this formulation of the problem, end-to-end architectures can be used, requiring less processing and taking better advantage of the amount of existing multimedia information.

Other widely used techniques such as Dynamic Time Warping (DTW) seek to directly compare the acoustic characteristics by constructing a frame-level similarity matrix [2]. The main idea of this method is to find the optimal warping path with the smallest distortion. DTW algorithm has been found to work best on high-level features, such as phone posteriorgrams [3]. However, these features are not available for low resource languages. Furthermore, the dynamic programming algorithm for similarity measure is time-consuming.

In this evaluation we apply a novel Two-Way attention mechanism also known as Attentive Pooling Networks. This method was successfully applied in the context of Query-by-example Spoken Term Detection in read speech in English and using word alignments [4]. Our objective is to analyze the performance of this approach under more realistic conditions, evaluating it on the scenario of natural speech in Spanish that is proposed in the Albayzin Search-on-Speech 2020 evaluation.

2. Attentive Pooling Networks for Query by Example

In this section, we present our primary system proposed for the Query-by-Example Spoken Term Detection (QbE-STD) task. This system is based on Attentive Pooling Networks, a recently proposed method for QbE-STD showing promising results [4].

2.1. System Description

From now on, we denote the spoken query, or more precisely the acoustic feature sequence, as $Q = \{q_1, q_2, ..., q_M\}$ and an audio segment where the query will be searched by $S = \{s_1, s_2, ..., s_N\}$, where M and N are the number of frames of the query and audio segment, respectively.

a. Embedding representation of audio segments

A Long Short-Term Memory (LSTM) network is used to obtain embedding representations of the two input audio segments. LSTMs are capable of storing temporal information, in theory, for long time spans [5]. Given a spoken query $Q = \{q_1, q_2, ..., q_M\}$, LSTM units project the query into a hidden state sequence $H_Q = \{h_1^Q, h_2^Q, ..., h_M^Q\}$, where h_M^Q contains information of the whole query. In the same way, the audio segment $S = \{s_1, s_2, ..., s_N\}$ is encoded into a second hidden state sequence $H_S = \{h_1^S, h_2^S, ..., h_N^S\}$. The same LSTM is used for input and query, so we expect that the representation of the query and the audio segment are in the same vector space.

b. Two-Way Attention: Attentive Pooling Networks

Now, we describe the Two-Way Attention mechanism, also called attentive pooling networks [4]. Figure 1 shows the structure of the system. First, the audio query $Q = \{q_1, q_2, ..., q_M\}$ and the audio segment $S = \{s_1, s_2, ..., s_N\}$ are encoded into the hidden vector sequences $H_Q = \{h_1^Q, h_2^Q, ..., h_M^Q\}$ and $H_S = \{h_1^S, h_2^S, ..., h_N^S\}$ by the shared RNNs to project the acoustic features into a vector space where they are more easily compared. Then the attention matrix G is computed as follows:

$$G = tanh(H_Q^T U H_S) \quad (1)$$

10.21437/IberSPEECH.2021-14

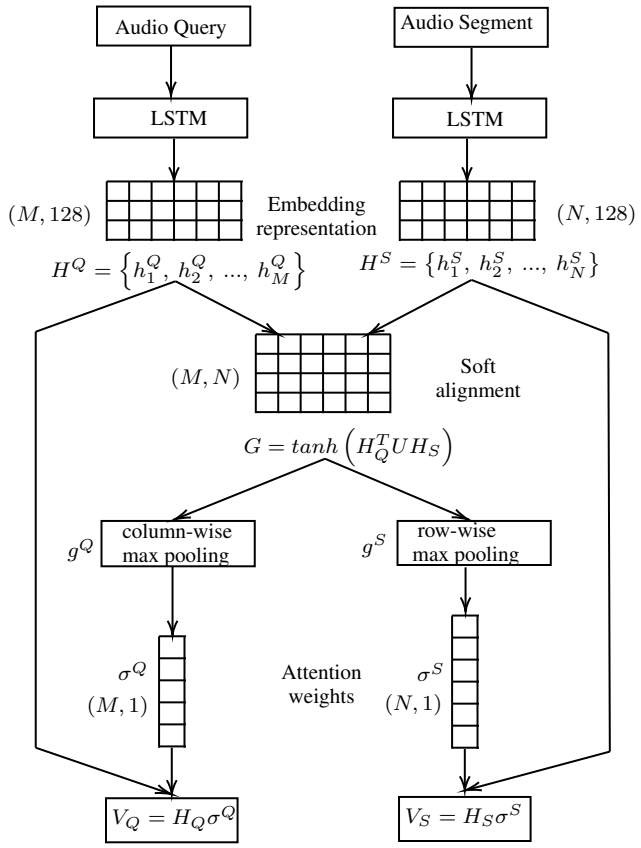


Figure 1: *Structure of Two-way attention (Attentive Pooling Networks) applied for QbE-STD in our Primary System*

This matrix is the main key to compare both, segment and query hidden state vector sequences. We can see matrix U as a measure of joint H_Q and H_S representations, and it is learned by training. To build a system more symmetric, we limit the measure matrix U to be a symmetric matrix, $U = U^T$. This allows to exchange the query and the segment without changing the result, because of the following relation.

$$H_Q^T U H_S = H_S^T U^T H_Q = H_S^T U H_Q \tag{2}$$

We initialize U with random normal samples with zero mean and 10^{-4} variance. Under these conditions, we can interpret G as a soft alignment score between each frame of the spoken query and the audio segment.

The next step is to apply a row-wise and column-wise max-pooling to the G matrix in order to generate the weight vectors $g^Q \in \mathbb{R}^M$ and $g^S \in \mathbb{R}^N$. This pooling is computed as follows.

$$[g^Q]_j = max_{1 \le i \le N}[G_{j,i}] \tag{3}$$

$$[g^S]_i = max_{1 \le j \le M}[G_{j,i}] \tag{4}$$

Note that the j-th element of vector g^Q represents the weight that we will apply to the j-th frame in the spoken query Q. Consequently, the result would be an estimation of which frames of the input sequences are actually important to make the comparison. Since that g^Q and g^S are attention vectors, it is necessary to normalize them with a softmax function to generate the final representation of attention vectors σ^Q and σ^S.

Finally, if we compute the dot product between the original hidden states and the attention vectors estimated, we have two representation vectors more comparable between them, because we have enhanced the parts that are really important to differentiate them. Vectors V_Q and V_S are computed as follows.

$$V_Q = H_Q\sigma^Q, V_S = H_S\sigma^S \tag{5}$$

c. Large-Margin Training

The whole system is trained using a large margin cost function (hinge loss). This is because the principal goal is to maximize the distance between classes and minimizing the intra class distance in each epoch. To achieve this, the training set is structured as groups formed by three elements: a spoken query $Q = \{q_1, q_2, ..., q_M\}$, a positive segment $S^{(P)} = \{s_1^{(P)}, s_2^{(P)}, ..., s_N^{(P)}\}$ and a negative segment $S^{(N)} = \{s_1^{(N)}, s_2^{(N)}, ..., s_N^{(N)}\}$. The positive segment contains the same word as the query, while negative segment is an aleatory audio segment taken from the training set that does not contain the same word as the query. Then, for each group we form the tuples $(Q, S^{(P)})$ and $(Q, S^{(N)})$ in order to calculate both of the attention matrices G for each tuple and the corresponding vector representations $(V_Q^{(P)}, V_S^{(P)})$ and $(V_Q^{(N)}, V_S^{(N)})$. Note that the V_Q vector representation is different depending on the audio segment for which it is computed, due to the Two-Way Attention. Cosine distance is used to measure how these tuples look alike. Cosine distance between V_Q an V_S is computed as follows.

$$l(V_Q, V_S) = (1 - cos(V_Q, V_S))/2 \tag{6}$$

We aim to minimize the distance between the query and the positive segment and to maximize distance between the query and the negative segment. For that a hinge objective function is used, defined as follows (M is the maximum possible distance, 1 in our case).

$$L_{hinge} = max\{0, M + l((V_Q^{(P)}, V_S^{(P)})) - l((V_Q^{(N)}, V_S^{(N)}))\} \tag{7}$$

3. Experimental Setup

As acoustic features we use 13 dimensional MFCC, which are extracted using the Kaldi toolkit [6] and Python Speech Features library with a sliding window with length of 0.025 seconds and a step of 0.01 seconds. These features are used in all the experiments proposed.

The shared RNNs consist of 2 layers with 128 LSTM units on all the models. Matrix U is a symmetric square matrix initialized with random normal values (with zero mean and a variance of 10^{-4}). As optimizer, we use Adam with a learning rate of 0.00005 and a minibatch of 128. We also use the WebRTC Voice Activity Detector (VAD) on all audio queries in order to remove all fragments of silence and random noise that could appear at the start and at the end of the audio recordings. All the neural networks are implemented with Pytorch and are trained for 4 epochs.

3.1. LibriSpeech Database description

a. LibriSpeech Training set

We use audio segments from the LibriSpeech Corpus [7]. We extract the segments from aligned utterances (with previously trained HMM models), so it is not necessary to use the VAD

detector because each segment contains a complete word exactly. Furthermore, we choose those segments that have at least 6 phonemes and a duration between 0.5 and 1.0 seconds. Then, we randomly select 500 different words and we form a training sample taking 2 segments containing the same word (one for the query (Q) and other for the positive segment (S^P)) and 1 segment containing a different word for the negative segment (S^N). For each training epoch we have a total of 1000 of such groups, consisting in (Q, S^P, S^N).

b. LibriSpeech Test Set

We use a separate LibriSpeech test set divided into two subsets. First, we define a set with 150 words that appear in the training set as queries. We call these IV (In-Vocabulary) segments. Second, we have another set with 150 words, but in this case none of this words appear in the training set as queries. We call these OOV (Out-Of-Vocabulary) segments.

For each word in the full test set, we randomly select 20 different words to compare with (IV and OOV segments). We force to have at least 5 words equal to the query to ensure a sufficient number of positive comparisons. Like in the training set, we force each word to have at least 6 phonemes and a duration between 0.5 and 1.0 seconds. Finally, the LibriSpeech test set is formed by 3000 IV segments and 3000 OOV audio segments.

Mean Average Precision (MAP), the mean of the average precision in the range of recall for each query in the testing set and P@20, the precision considering the 20 best scores, are the evaluation metrics employed for LibriSpeech test Set.

3.2. Albayzin Search on Speech Database description

a. Development data

Training and development data provided by the evaluation organizers belong to the MAVIR and RTVE databases.

- For MAVIR database, development list of terms have about 375 different terms (375 OOV for our systems) whose length ranges from 5 to 27 single graphemes for Spoken Term Detection task. There is a total of 100 queries with three examples per query. Some of the queries have a fixed length of 3 seconds of audio, including a large amount of silence. We used webrtcVAD on all the segments to remove the silence before using them.
- For RTVE database, two different datasets are provided: *dev1* and *dev2*. Dataset *dev1* is not used in this paper. The dataset *dev2* consists of about 400 different terms (all are OOV for us) whose length ranges from 4 to 25 single graphemes. Like in MAVIR database, there is a total of 100 queries with three examples per query. We used webrtcVAD on the RTVE queries to reduce the large amount of silence appearing in some of them.

b. Test data

Three databases are provided by the organizers for system evaluation:

- MAVIR test speech data consists of about 200 different terms whose lengths range from 4 to 28 single graphemes. For the Query-by-Example Spoken Term Detection task, the systems implemented are tested with about 100 queries.
- RTVE test data consists of about 400 different terms of the RTVE program material given by the evaluation. RTVE test data has a total of about 400 different terms whose length ranges from 4 to 28 single graphemes. Again, this data set is composed by 100 queries and three examples per query.
- SPARL20 test data consists of a subset of Spanish parliament sessions. SPARL20 has a total of 200 different terms whose length ranges from 3 to 19 single graphemes. Like in the other testing sets, this set is conformed by 100 queries with three examples per query.

ATWV (Actual Term Weighted Value) and MTWV (Maximum Term Weighted Value) are the metrics proposed by the evaluation. We use the VAD detector webRTCVAD on all the test data queries.

4. Results on LibriSpeech test set

In this section we describe two different experiments performed and their corresponding results when evaluated on the LibriSpeech test Set. Both experiments differ in the alignment of the segment used to search for the query.

In experiment 1, we train the Two-Way attention model with LibriSpeech Training set without introducing noise and using the word alignments to train and evaluate the system. This evaluation is highly unrealistic since it requires to have the word alignments of the segments on which the query is searched. For that reason we consider an alternative setup.

In experiment 2, we aim to evaluate the performance of the system trying to simulate the more realistic setup of not having the word alignments of the segments on which the query is searched. We start from the same set used in experiment 1 and, for each pair (Q, S) we look for two random words from the entire subset not coincident with the words contained in Q and S. Then, we concatenate these words to the beginning and end of the segment, and trim the segment to a fixed length of N = 150 frames. In this way the segment does not contain a perfectly aligned word. This technique has been applied both for training and testing, so that this experiment simulates the possibility of not having word alingments in both training and test.

Table 1: *Performance of QBE-STD on LibriSpeech Testing Set*

Exp.	MAP (IVs)	MAP (OOVs)	MAP (Total)	P@20
Exp.1	0.981	0.971	**0.976**	**0.23**
Exp.2	0.972	0.953	0.962	0.06

As can be seen in the Table 1, better results are always obtained in experiment 1. As expected, results worsens when the word alignments are not available. However, the results obtained without word alignments (exp. 2) are still very good, since the total MAP is very close to one, indicating that the system is committing very few false positives. When considering the precision for the 20-best scores, the degradation in performance becomes more clear. Our results are much better than those reported in the original paper proposing this approach [4], most probably because testing was not performed in the same conditions. In particular, we have forced 5 of the 20 segments compared against each query to contain the same word as the query, while this was not enforced in the original paper.

5. Albayzin Search-on-Speech 2020 QbE-STD systems and results

This section describes the systems submitted to the 2020 edition of ALBAYZIN Search-on-Speech Query-by-Example Spoken Term Detection (QbE-STD) evaluation and the results obtained with them.

5.1. System Description

All of the systems submitted are based on the Two-Way attention mechanism (Attentive Pooling Networks) and trained on LibriSpeech (an English database). In some of the systems development data in Spanish has been used to try to adapt the system to the language of the evaluation. For each system, threshold was set to make ATWV as close as possible to MTWV on development data.

a. Primary System

For the primary system we have used the neural networks trained in LibriSpeech experiment 1, since we have not found improvements on development data by adding spoken word segments to the beginning and end of the segment to be compared (experiment 2). For this primary system, we have retrained the neural networks for two epochs with the MAVIR development set. The goal was to apply transfer learning [8] to adapt the system trained on a large database in English with a small database in Spanish.

As explained in the experimental setup, we first apply the WebRTC Voice Activity Detector to the 100 queries and then we go through the entire segments to be analyzed with an adaptive sliding window with a length dependent on the length of the query, since we expect the ocurrences of the query to be of similar size to the query.

b. Contrastive System 1

In this case, we have used the neural networks trained from LibriSpeech experiment 1 again, this time without modifying the weights with data in Spanish. We have also used an adaptive window length as in the primary system.

c. Contrastive System 2

Contrastive system 2 is almost identical to contrastive system 1, they only differ in that in this case we are applying a z-score normalization approach using

$$z = \frac{x - \mu}{\sigma} \tag{8}$$

Where μ is the mean of the scores obtained and σ is the standard deviation of these. We wanted to evaluate if there was a score normalization issue with this contrastive system.

5.2. Results on Albayzin Search-on-Speech 2020 data

Table 2 presents development and test results obtained on the evaluation data. The first conclusion is that we have not been able to successfully apply the Two-Way attention mechanism (or Attentive Pooling networks) on the evaluation data. All results present very low and even negatives ATWVs. Even without taking into account threshold setting (looking at MTWV) results are poor. There seem to be a number of unresolved issues in the transition from a read speech scenario in English (LibriSpeech) with word-aligned data to a natural speech scenario in Spanish (MAVIR, RTVE and SPARL20) without word-aligned data that have produced these results.

As can be seen in the table 2, better results have generally been obtained with the primary system, which indicates that there is some improvement when retraining for a few epochs with audios in Spanish.

In the two contrastive systems, there is an obvious problem in setting the detection threshold, as the MTWV and ATWV values differ greatly. Score normalization did not provide consistent improvements either, as can be seen in the results for the contrastive 2 system.

Table 2: *Performance of QBE-STD Primary System (PRI), Contrastive System 1 (CON1) and Contrastive System 2 (CON2) on Albayzin2020 development and test data*

Dataset	System	MTWV	ATWV
MAVIR DEV	PRI	**0.0533**	**0.0491**
	CON1	0.0160	-38.5775
	CON2	0.0000	-158.2873
MAVIR TEST	PRI	**0.0126**	**-0.1061**
	CON1	0.0000	-393.5610
	CON2	0.0000	-38.5959
RTVE DEV	PRI	**0.0465**	**0.0465**
	CON1	0.0414	-76.0473
	CON2	0.0000	-51.9993
RTVE TEST	PRI	**0.0209**	-115.7086
	CON1	**0.0209**	-88.3716
	CON2	0.0000	**-16.5831**
SPARL20 TEST	PRI	0.0107	**0.0107**
	CON1	**0.0306**	-34.2099
	CON2	0.0000	-103.6805

6. Conclusions

In this evaluation we have tried to apply a novel Two-Way attention mechanism also known as Attentive Pooling Networks. This approach was successfully applied in the context of Query-by-Example Spoken Term Detection in read speech in English (LibriSpeech) and using word alignments [4]. We have been able to reproduce good results on this scenario and have tried to improve the system by simulating the more realistic scenario of not having word alignments. However, this approach has not helped when transitioning from this scenario to the scenario of natural speech in Spanish that is proposed in the Albayzin Search-on-Speech 2020 evaluation. We have obtained limited improvements when retraining the system on a small amount of Spanish data but, in the end, the results obtained on the evaluation data are poor compared to other more classic alternatives. Anyway, we still consider that this approach has the potential to compete in results with these more classical approaches, and we will continue exploring it in the future.

7. Acknowledgements

Work developed under project DSForSec (RTI2018-098091-BI00), funded by the Ministry of Science, Innovation and Universities of Spain and FEDER.

8. References

[1] D. R. Miller, M. Kleber, C. L. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," *International Speech Communication Association - 8th Annual Conference of the International Speech Communication Association, Interspeech 2007*, vol. 3, pp. 1965–1968, 2007.

[2] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[3] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," *Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2009*, pp. 421–426, 2009.

[4] K. Zhang, Z. Wu, J. Jia, H. Meng, and B. Song, "Query-by-example spoken term detection using attentive pooling networks," *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2019*, pp. 1267–1272, 2019.

[5] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2015-August, pp. 5206–5210, 2015.

[8] S. Panigrahi, A. Nanda, and T. Swarnkar, "A Survey on Transfer Learning," *Smart Innovation, Systems and Technologies*, vol. 194, no. 10, pp. 781–789, 2010.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

GTH-UPM System for Albayzin Multimodal Diarization Challenge 2020

Cristina Luna-Jiménez[1], *Ricardo Kleinlein*[1], *Fernando Fernández-Martínez*[1], *José Manuel Pardo-Muñoz*[1], *José Manuel Moya-Fernández*[1]

[1]Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid, 28040, Madrid, Spain

cristina.lunaj@upm.es, ricardo.kleinlein@upm.es, fernando.fernandezm@upm.es, josemanuel.pardom@upm.es, jm.moya@upm.es

Abstract

This paper describes the multimodal diarization system proposed by the GTH-UPM team to Albayzin Multimodal Diarization Challenge 2020. The submitted solution consists of 2 separate diarization systems that work on visual and aural components.

The visual diarization solution exploits web resources, as well as provided enrollment images. First, these images feed a facial detector. Next, all the discovered faces are introduced into FaceNet to generate embeddings. After this, we apply a clustering algorithm on extracted embeddings, obtaining a representative cluster for each participant. Each centroid of the representative clusters acts as a participant model. When a new embedding extracted from a facial image of the program arrives at the system, it receives the label that corresponds to the closest centroid identity among all the given participants, as long as it exceeds a fixed quality threshold.

The aural speaker diarization problem is tackled as a classification task, in which a deep learning model learns the mapping between automatically-extracted sequences of aural x-vectors and speaker identities. These sequences aid in overcoming the scarcity of training samples per speaker.

The best results sent reached a DER of 66.94% for visual diarization and a DER of 125.24% for aural diarization on the test set.

Index Terms: Multimodal diarization, clustering, biLSTM, attention

1. Introduction

Speaker diarization systems aim to identify who is talking and when [1]. Similar to speaker diarization, visual diarization systems try to discover who is in the scene and when. Although visual diarization is not as common as speaker diarization, there are already some publications that start to include this modality to reduce the diarization error rate [2][3]. Evidence of this tendency change is the Albayzin Multimodal Diarization Challenge at IberSPEECH conference [4].

Previous year approaches [5][6][7] address visual and aural diarization in different ways, but all of them include some common modules. These modules consist of a tracker and a detector of participants or speech, a feature extractor to generate embeddings, and an identity recognizer. In line with prior work, we propose a visual diarization system with mentioned modules, including information extracted from a novel source, Google Images.

Although traditionally speaker diarization has been tackled from an unsupervised point of view [8, 1], recent research seems to successfully approach the problem from a supervised perspective by means of recurrent neural models [9]. Additionally, attention mechanisms have lately enabled great improvements in many fields, including speaker diarization [10]. On a similar basis, x-vectors were proposed by Snyder *et al.* as a way to map variable-length utterances to fixed-dimensional embeddings that greatly enhance speaker's acoustic characterization [11]. Therefore, we propose a speaker diarization system based on the supervised learning of a map between fixed-length sequences of x-vectors and speaker identities via recurrent and attention neural models.

The structure of the paper is as follows: Section 2 describes the proposed diarization system. Section 3 summarises the computational cost of the experiments. Section 4 presents the main experiments performed and the results obtained for development and test sets. Finally, in Section 5, we illustrate the extracted conclusions and future research work.

2. System description

In this section, we present the visual and aural diarization systems. The visual diarization system bases its functionality on a weakly-supervised strategy, solving the diarization and the attribution tasks. Regarding speaker diarization, we employ a fully-supervised model, trained on the development programs.

In what follows, we explain the different components that constitute them and their relationships to detect who is speaking or appearing in each program.

2.1. Visual Diarization Pipeline

Discerning identities from visual information still presents several challenges intrinsic to videos, like dealing with occlusions, blurring, etc. Our proposed solution overcome some of these challenges through four main modules: data acquisition, facial detection, identity recognition, and post-processing. A whole picture of the pipeline is in Fig. 1.

To homogenize the programs and accelerate the experiments, we worked at 5 fps with a resolution of 1.024 x 576.

2.1.1. Data acquisition

Due to the limited amount of images provided in the enrollment set, we included data acquired from Google Images. To afford it, we developed a tool to download these resources automatically. One inconvenience of this methodology is the uncertainty about the picture's content, i.e. whether downloaded images belong to the queried identity. To deal with noisy images and discard them, we apply a filtering process. This process consists of applying a face detector and a face recognizer over all the available material.

10.21437/IberSPEECH.2021-15

2.1.2. Face detection

Both images downloaded from Google and those provided in the enrollment set are passed through MTCNN [12]. MTCNN is a state-of-the-art face detector that returns face positions and their probability of being faces. In our settings, we fixed acceptance confidence of 98% and minimum face size of 80x80. Detected faces that do not accomplish these thresholds are removed.

2.1.3. Face recognition

To perform facial recognition, firstly, we extract embeddings from a pre-trained network, secondly, we make a clustering with these embeddings and, finally, we build a model per participant based on clustering results.

For the embeddings generation, we re-scale detected faces by MTCNN to a size of 160x160. These resized facial images are injected into FaceNet [13]. FaceNet is a state-of-the-art pre-trained model in identity recognition. From each facial image introduced, FaceNet returns a 128-dimensional vector. This embedding represents the face in a latent space. Face embeddings corresponding to the same identity are closer to each other in the latent space than those corresponding to different identities.

As commented before, Google Images could return unexpected results. To remove the non-useful material, we apply a clustering algorithm. DBSCAN [14], from sklearn library [15], is the clustering algorithm selected since it does not require to set the desired number of clusters beforehand.

To detect the optimal working point of the participant clustering, we run several DBSCAN [14] instances, varying the epsilon value. More specifically, we scan epsilons between 1 and 12, in steps of 0.5. The rest of the DBSCAN parameters maintain their default value, except for the min_samples, fixed to 5. Once the scanning finishes, the optimal instance agrees with the maximum silhouette [16] coefficient obtained in the scanning. The silhouette coefficient is an unsupervised metric that measures the quality of each DBSCAN instance in terms of inter and intra-cluster distances.

Having the optimal DBSCAN [14] instance, we obtain the representative cluster of each identity by using this score:

$$SC(E_{CLT_i}, E_{ENR_j}) = w * E_{ENR_j} + (1 - w) * E_{CLT_i} \quad (1)$$

where E_{CLT_i} represents the percentage of embeddings in cluster i over the total number of embeddings of the participant j; E_{ENR_j} is the percentage of images of the enrollment set of participant j that belong to cluster i; and w is the weight that balance the contribution of each term.

By varying the equation's weight, we can increase or reduce the contribution of the enrollment images. A weight equal to 0 selects the cluster with the maximum number of embeddings as the representative. However, with a weight equal to 1, the algorithm chooses the cluster that contains more embeddings of the enrollment set as the representative, independently if it is the densest or not.

As the competition provides the enrollment images, they probably contain the face of the participant and can be used as a reference to know whether the cluster is of the expected participant or not. If the used images would be the images obtained from Google exclusively, the predominant cluster might not coincide with the expected participant since Google is noisy. For this reason, we compare the performance of the algorithm with two weights: 0 and 0.9 to study the effect of using the enrollment images as a guide against not using it, relying only on Google data. See Table 1.

We repeat this process with each participant. Thus, for each participant, we discard wrong downloaded pictures and select his/her representative cluster.

Finally, the character's model of each participant is the centroid of the representative cluster, i.e. the average vector given by all the embeddings that constitute its representative cluster.

2.1.4. Post-Processing

Intending to improve results, we included in the pipeline several post-processing techniques to refine results.

When a new face arrives at the system, the system calculates the cosine similarity between the embedding of the face and the key-participants models, i.e. the centroid of the representative cluster. The new face receives the label of the closest key-participant centroid, i.e. the key-participant with the highest cosine similarity. To reduce the false-positive rate, we fixed a quality threshold of 0.5. This quality filter lets modify the label from a key-participant to 'unknown' whenever its cosine similarity is lower than 0.5. Check Prediction & Evaluation box in Fig. 1.

To deal with occlusions and blurring in some frames, we applied a tracking process to extend predictions from a single frame to a track. We apply tracking in 2 steps:

1. Detect scenes: FFMPEG implements a scene detector. This detector applies the sum of absolute differences and returns a probability of scene change in each frame. When the frames' probability overpasses a certain threshold, it marks the start of a new scene. In the experiments, we use this detector with a threshold of 0.2.
2. Full-body tracking: we track people in each scene using a multi-person tracker [17]. This repository makes use of several pre-trained classification models able to detect people on an image. In our case, we selected the YOLO model and a people detection threshold of 0.7. The algorithm performs tracking by distinguishing overlapping bounding boxes in consecutive frames.

To match full-body bounding boxes, obtained in tracking, with facial bounding boxes, detected by MTCNN [12], we calculate the intersection over union (IoU) of all the pairs. The pair of full-body and facial bounding boxes that reaches the maximum IoU is matched. Once, there is a match between the face and the tracked body, it is possible to assign a single identity to the whole track, correcting face miss-detections and possible wrong predictions. The assigned identity to the whole track is the maximum voted, or the most times predicted along the faces that conform the track.

2.1.5. Prediction of identity

By passing new frames through all the modules, we extract the diarization file with all the participants' predicted in each moment. From this result, we generate a file called 'face pooling', that contains the list of participants detected by the visual diarization system at least once during the program. This reduced list of participants is a new source of information that the aural diarization applies to fine-tune the model.

2.2. Aural Diarization Pipeline

Speaker diarization systems aim at identifying who is talking and the timestamps that define such talking turn. Our proposed method treats the problem in the first stage as a supervised classification task, matching chunks of x-vectors with speaker

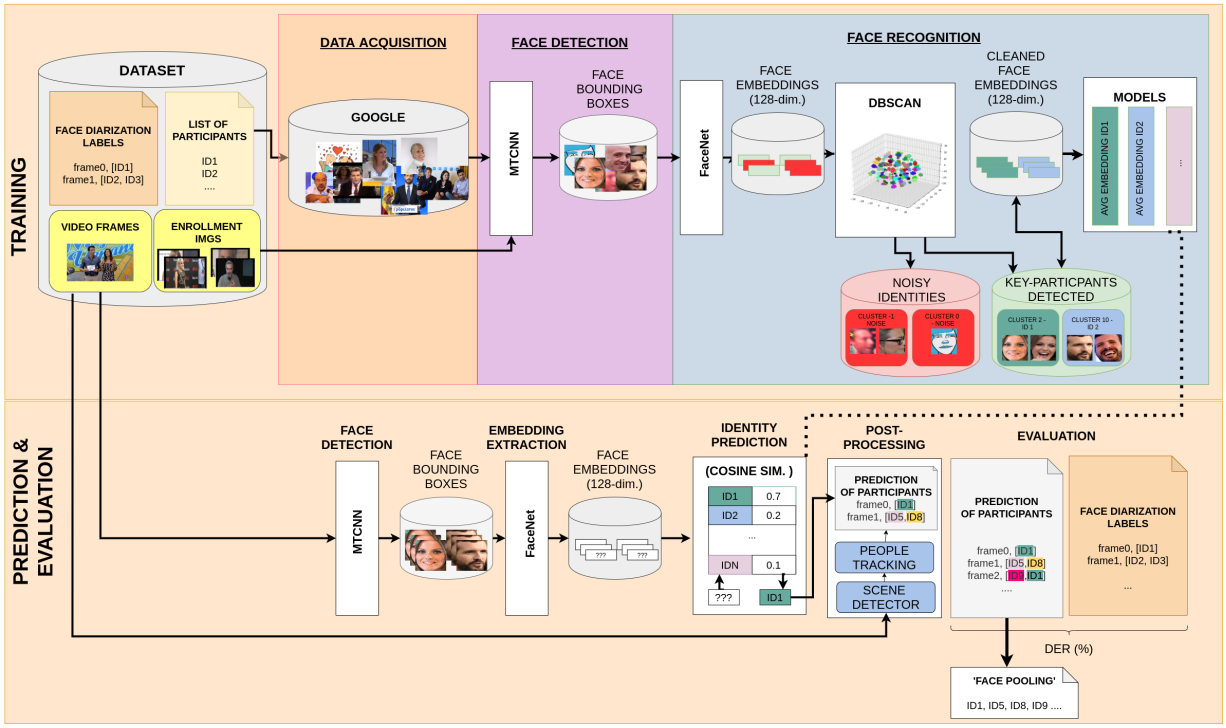


Figure 1: *Visual diarization pipeline*

names on the audios provided in the development set and the chunks of the enrollement set. Secondly, the temporal alignment of the samples let turn a classification problem into a speaker diarization estimation.

2.2.1. *Extraction of aural embeddings*

First of all, audio files originally provided in AAC format are converted into 16KHz, single-channel WAV format using the FFMPEG toolbox [18]. Next, we compute 512-dimensional aural x-vectors for every audio file following the default procedure explained in the III DIHARD competition [19]. We extract x-vectors every 0.75 seconds, spanning 1.5 seconds with 50% overlap between adjacent vectors. This procedure is applied over development, enrollment, and test data.

2.2.2. *Speaker Classification*

We desire to estimate a speaker name for every x-vector passed as input to the system. However, rather than a single x-vector, the predictive model's input is extended to include 5 consecutive samples. A sliding window extracts these samples by taking the current x-vector and its near context. Despite receiving a sequence of x-vectors as input, the outcome is a single prediction. This prediction is associated with the x-vector that occupies the middle position of the input sequence.

The classifier's architecture has an initial biLSTM layer with 50 units, followed by a self-attention layer and a fully-connected layer. In the end, a softmax activation layer provides speaker probabilities.

The classifier minimizes the cross-entropy loss using an Adam optimizer [20]. Both biLSTM and the attention layers have a dropout rate of 0.3. The initial learning rate is 0.001. We also include an early-stopping regularization that stops the model's training after 5 epochs without increasing the F1 classification score. To increase the robustness of the model from epoch to epoch, we implement a data augmentation technique. This approach adds white noise to the input embedding sequences at a random probability rate of 0.15.

For the system submitted as primary (p in Table 1), the learning process consists of two folds: First, we teach the base model (c1 in Table 1) to classify among all the potential identities given in development, test, and enrollment data, until convergence.

For training this model, we use the audios provided in the development set, that consist of the programs' records and the clips of the enrollment folder. To include the participants' voices that appear in the test set but not in the development set, we also add the clips of the test enrollment folder.

Once the base model is trained, we adapt it to each program by applying fine-tuning with the identities in the "face pooling" returned by the visual diarization branch.

To fine-tune the model, we get all the embeddings in the training set that include a participant of the 'face pooling' and re-train the model with this reduced set of participants, maintaining the previously mentioned configuration.

Therefore we end up having a fine-tuned and specialized model for every test program.

2.2.3. *Speech/Non-speech Segmentation*

Voice Activity Detectors (VADs) are a customary solution to discern non-speech parts of the audio from actual speech. We employ the default DIHARD's VAD module over the test partition data [19]. This step enables to remove the parts predicted as non-speech. This filtering allows refining our predictions to a greater degree of detail.

In environments such as TV broadcast debates, in which

talking turns can change in a matter of milliseconds due to interruptions, the use of a VAD-based speech filter gains importance.

3. Computational Cost

Visual diarization experiments were carried out in a computer with an Intel® Core™ i7-3770K Processor, 32GB RAM and a GPU NVIDIA Titan X Pascal, 12 GB. The most time-consuming tasks were the FaceNet embeddings extraction and the tracks calculation. In total, the version without tracking and no-parallel processing takes $xRT \approx 1.18$.

Regarding the aural component, the overall expense in time account up to 2 hours and a half approximately. We used a Nvidia GeForce RTX 2070 with 8Gb of RAM memory. The rest of computer specifications are the same as described for the visual part.

4. Experiments and Results

Table 1 collects the best results obtained in each modality. The best system submitted to the competition in the visual modality for the development set (the second row in Table 1) reached a DER of 66.94% on the test set. This system performs cluster assignment with a weight of 0.9, giving more relevance to the number of enrollment images. Additionally, it includes tracking as post-processing.

The first system in Table 1 is the same as the previously mentioned without including the tracking module. As we can see, the tracking strategy seems to improve the final Diarization Error Rate in the development set, but not in the test set. This mismatch-effect is explained by the different types of programs that define the test set, an effect that our tracking solution can not address since it uses the same parameters for all the programs. To confirm this behavior, we repeated the experiments adapting the tracking parameters to the test set, and although the DER of the test set decreased (60.32%), the DER in the development set increased (73.93%).

After analyzing the results per program for the tracking configuration of Table 1 against the non-tracking version, we detect that DER in the test set for programs with acronym AT, BR, NFMY, and WU decreases in 4.97%, 7,42%, 1.24%, and 11.20%, respectively, using tracking module; meanwhile for programs BN, CA, EP, LD, ML, and SFT the DER increases in 5.55%, 38.74%, 1.04%, 4.41%, 157.2%, and 9.42%, respectively, when we use the tracking module. Notice that these values are the average per type of program, not considering durations.

We performed additional experiments not shown in Table 1, using only Google Images (DER = 76.55%) and using only enrollment images (DER = 75.41%) with w=0.0 and without tracking. Experiments reveal that joining both sources of information can improve diarization results (DER = 75.24 %), although the difference is not too significant since the model uses the average of facial embeddings that softens the effect of having more images.

Regarding aural diarization, it seems that the system based on 'face pooling' performs worse than the version without fine-tuning the base model. Uniquely, programs of type BN, CA, NFMY, SFT, and WU improves DER values. These results reveal that in spite of the reduction in the number of input identities to the model (40%), it is not enough to compensate the 7% of lost identities. Furthermore, it is essential to mention that the model employed in speaker diarization is fully-supervised, which means that it is tailored to the training domain. As a result, when it faces different programs or lack of data of some participants, it reduces its performance, as has happened in this challenge.

Table 1: *Diarization Error Rate for visual and aural systems*

System	DER dev. (18 IDs)	DER test (161 IDs)	Run
1. Visual Diarization w=0.9 - No Tracking	74.34 %	61.58 %	-
2. Visual Diarization w=0.9 - Tracking	**59.63 %**	**66.94 %**	**p**
3. Aural Diarization without 'face pooling'	**-**	**125.24 %**	**c1**
4. Aural Diarization with 'face pooling'	-	131.59 %	p

5. Conclusion and future work

In this paper, we present the GTH-UPM systems employed for solving the 2020 Albayzin Multimodal Diarization Challenge. The visual recognition solution is based in a weakly-supervised strategy that employs web resources, clustering and distances to obtain the participants that appear in each scene. Additionally, it also implements tracking and post-processing techniques to improve overall performance. To combine aural and visual diarization, we extract the participants detected in the visual diarization to fine-tune a biLSTM model with an attention mechanism.

Although results encourage us to continue with this research line, there are still some open issues to address in the future. One of the most important is how to exploit the images acquired from Google to improve visual diarization. We plan to test a more complex model rather than clustering to perform facial recognition and enhance the tracking module to automatically adapt its parameters to the type of program.

Regarding aural speaker diarization, we plan to explore new alternatives to increase the amount of available data to train the classifier model. Among such, we contemplate incorporating additional data augmentation techniques or, as we did in the visual diarization pipeline, recovering material from a second source of information.

Concerning fusion, we also consider investing some effort in testing other strategies to improve global DER combining embeddings of both modalities.

To conclude, this paper contributes to web resources' exploitation and the study of diarization systems of two (visual and aural) modalities.

6. Acknowledgements

The work leading to these results has been supported by the Spanish Ministry of Economy, Industry and Competitiveness through the CAVIAR (MINECO, TEC2017-84593-C2-1-R) and AMIC (MINECO, TIN2017-85854-C4-4-R) projects (AEI/FEDER, UE).

Ricardo Kleinlein's research was supported by the Spanish Ministry of Education (FPI grant PRE2018-083225).

We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

7. References

[1] I. Viñals, A. Ortega, J. Villalba, A. Miguel, and E. Lleida, "Domain adaptation of PLDA models in broadcast diarization by means of unsupervised speaker clustering," in *Proc. Interspeech 2017*, 2017, pp. 2829–2833. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-84

[2] P. Campr, M. Kunešová, J. Vaněk, J. Čech, and J. Psutka, "Audio-video speaker diarization for unsupervised speaker and face model creation," in *Text, Speech and Dialogue*, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds. Cham: Springer International Publishing, 2014, pp. 465–472.

[3] P. A. Marín-Reyes, J. Lorenzo-Navarro, M. Castrillón-Santana, and E. Sánchez-Nielsen, "Who is really talking? a visual-based speaker diarization strategy," in *Computer Aided Systems Theory – EUROCAST 2017*, R. Moreno-Díaz, F. Pichler, and A. Quesada-Arencibia, Eds. Cham: Springer International Publishing, 2018, pp. 322–329.

[4] J. Luque, A. Bonafonte, F. A. Pujol, and A. J. S. Teixeira, Eds., *Fourth International Conference, IberSPEECH 2018, Barcelona, Spain, 21-23 November 2018, Proceedings*. ISCA, 2018. [Online]. Available: https://doi.org/10.21437/IberSPEECH.2018

[5] B. Maurice, H. Bredin, R. Yin, J. Patino, H. Delgado, C. Barras, N. W. D. Evans, and C. Guinaudeau, "ODESSA/PLUMCOT at Albayzin Multimodal Diarization Challenge 2018," in *Fourth International Conference, IberSPEECH 2018, Barcelona, Spain, 21-23 November 2018, Proceedings*, J. Luque, A. Bonafonte, F. A. Pujol, and A. J. S. Teixeira, Eds. ISCA, 2018, pp. 194–198. [Online]. Available: https://doi.org/10.21437/IberSPEECH.2018-39

[6] M. A. I. Massana, I. Sagastiberri, P. Palau, E. Sayrol, J. R. Morros, and J. Hernando, "UPC multimodal speaker diarization system for the 2018 Albayzin Challenge," in *Fourth International Conference, IberSPEECH 2018, Barcelona, Spain, 21-23 November 2018, Proceedings*, J. Luque, A. Bonafonte, F. A. Pujol, and A. J. S. Teixeira, Eds. ISCA, 2018, pp. 199–203. [Online]. Available: https://doi.org/10.21437/IberSPEECH.2018-40

[7] E. Ramos-Muguerza, L. D. Fernández, and J. L. Alba-Castro, "The GTM-UVIGO system for audiovisual diarization," in *Fourth International Conference, IberSPEECH 2018, Barcelona, Spain, 21-23 November 2018, Proceedings*, J. Luque, A. Bonafonte, F. A. Pujol, and A. J. S. Teixeira, Eds. ISCA, 2018, pp. 204–207. [Online]. Available: https://doi.org/10.21437/IberSPEECH.2018-41

[8] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, 2003, pp. 411–416.

[9] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6301–6305, 2019.

[10] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 296–303, 2019.

[11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[12] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, p. 1499–1503, Oct 2016. [Online]. Available: http://dx.doi.org/10.1109/LSP.2016.2603342

[13] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2015.7298682

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[16] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, p. 53–65, Nov. 1987. [Online]. Available: https://doi.org/10.1016/0377-0427(87)90125-7

[17] M. Kocabas and C. Heinrich, "Simple multi person tracker," 12 2019. [Online]. Available: https://github.com/mkocabas/multi-person-tracker

[18] S. Tomar, "Converting video formats with ffmpeg," *Linux Journal*, vol. 2006, p. 10, 2006.

[19] N. Ryant, K. W. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "Third DIHARD Challenge Evaluation Plan," *ArXiv*, vol. abs/2006.05815, 2020. [Online]. Available: https://arxiv.org/pdf/2006.05815.pdf

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

ViVoLAB Multimodal Diarization System for RTVE 2020 Challenge

Victoria Mingote, Ignacio Viñals, Pablo Gimeno, Antonio Miguel, Alfonso Ortega, Eduardo Lleida

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain
{vmingote, ivinalsb, pablogj, amiguel, ortega, lleida}@unizar.es

Abstract

This paper describes a post-evaluation analysis of the system developed by ViVoLAB research group for the IberSPEECH-RTVE 2020 Multimodal Diarization (MD) Challenge. This challenge is focused on the study of multimodal systems for the diarization of audiovisual files and the assignment of an identity to each segment. In this work, we have implemented two different subsystems to address this task using the images and the audio from files separately. To develop our subsystems, we have employed the state of the art speaker and face verification embeddings extracted from publicly available Deep Neural Networks (DNN). Different clustering approaches are also used in combination with the tracking and identity assignment process. Furthermore, in the face verification system, we have included a novel approach to train an enrollment model for each identity which we have shown previously to improve the results compared to the average of the enrollment data. Using this approach, we train a learnable vector to represent each enrollment character.

Index Terms: face recognition, speaker recognition, deep neural networks, enrollment models, spectral clustering, video processing

1. Introduction

Multimodal biometric verification field consists of identifying people using audiovisual resources. In recent years, this field has been widely investigated due to its great interest which is motivated by the fact that human perception uses not only acoustic information but also visual information to reduce speech uncertainty. Furthermore, this task had been rarely addressed for uncontrolled data due to the lack of this kind of datasets. However, in recent years, several challenges focused on this topic have been developed [1, 2, 3], and also a large amount of multimedia and broadcast data is being produced currently like news, talk shows, debates or series. Therefore, to develop a multimodal biometric system, different tools are required to process these data, detect the presence of people and address the identification of who is appearing and speaking. The approach employed in this kind of systems is known as multimodal diarization.

Many studies focus on the simplest way to perform the multimodal diarization based on having separate systems for speaker and face diarization [3, 4]. Speaker diarization is a widespread task [5, 6] due to its usefulness as pre-processing for other speaker tasks. At the same time, it is still a challenging task because there is no prior information about the number and the identity of speakers in the audio files, and the domain mismatch between different scenarios can produce some difficulties. On the other hand, face diarization has been widely employed as a video indexing tool, and the previous step for face verification [7, 8]. However, in unconstrained videos of real-world scenarios, face images often can appear with large variations, so this kind of system has also found some problems in real-world scenarios. For these reasons, a straightforward score level fusion is usually employed to join the information of both types of systems.

The IberSPEECH-RTVE 2020 Challenges aims to benchmark and further analyze this different kind of diarization systems. With this purpose, two types of diarization evaluations are included in this challenge, Speaker Diarization and Identity Assignment (SDIA) [9], and a Multimodal Diarization (MD) [10]. The former is the most extended kind of diarization combined with the speaker assignment, while the latter combines the previous one with face diarization, which is obtaining more relevance in recent times. Thus, we have focused on this second challenge, and specially we will remark the characteristics of face diarization subsystem.

This paper presents the ViVoLAB system submitted to the IberSPEECH-RTVE 2020 Challenge in MD task. This challenge is focused on segmenting broadcast audiovisual documents and assigning to the segments an identity from a closed set of different faces and speakers. For the challenge, we have processed video and audio tracks independently in order to separately improve their performance. However, the pipeline is very similar in both cases where the differences are the exact approach used in each part of the process. Therefore, initially, the video and audio files are processed. After that, an embedding extractor is used to extract the representations, and finally, clustering and assignment process is applied. To carry out the assignment process in the face subsystem, a new approach based on [11] has been applied to model the enrollment identities. This approach was shown as a promising technique to characterize each enrollment identity with only one learnable vector for the speaker verification task, but this is the first time that this technique has been applied in face verification.

The remainder of this paper is laid out as follows. Section 2 provides a description of the challenge and the dataset employed. In Section 3, we describe the face diarization subsystem. The speaker diarization employed is explained in Section 4. Finally, Section 5 presents and discusses results, and Section 6 concludes the paper.

2. RTVE 2020 Challenge

The RTVE 2020 Challenge is part of the 2020 edition of the Albayzin evaluations [12, 10]. This dataset is a collection of several broadcast TV shows in Spanish language and covering different scenarios. To carry out this challenge, the database provides around 40 hours of shows from the public Spanish Television (RTVE). The development subset of the RTVE2020 database contains two of the parts of the RTVE 2018 database ($dev2$ and $test$ partitions) which are formed by four shows of around 6 hours. Furthermore, this subset also contains a new development partition with nine shows of around 4 hours. The evaluation set consists of fifty-four video files of around 29 hours in total with speaker and face timestamps. Enrollment

10.21437/IberSPEECH.2021-16

data is also provided for 161 characters with 10 pictures and a 20-second video of each character.

3. Face Subsystem

This section describes the different blocks of the face system, including video processing, embedding extraction, training face enrollment models, clustering, tracking, and identity assignment scoring. The block diagram of the face system is depicted in Fig.1.

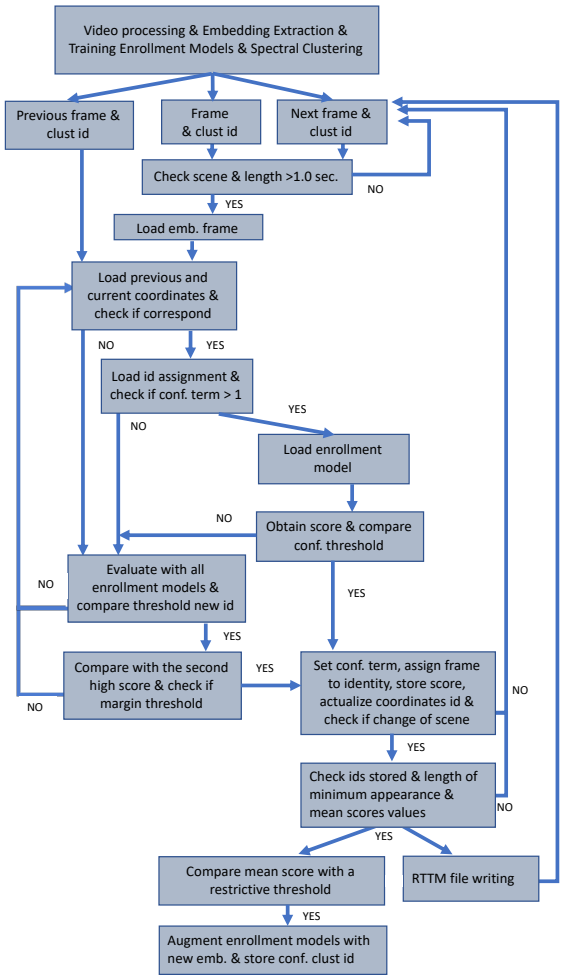


Figure 1: *Block diagram of face system.*

3.1. Video Processing

3.1.1. Frame Extraction

As the first step, we process the video to extract five frames per second using $ffmpeg$ tool [1]. We decided to use five frames per second since this number of frames allows us to have a high precision to determine the limits of the characters appearance.

3.1.2. Face Detector

Face detection is a fundamental step because failures in this process could be crucial for correct development in other parts of the face diarization system. In our system, the face detector employed is a system of alignment and detection based on a deep neural network (DNN) which is called Multi-task Cascaded Convolutional Networks (MTCCN) [13]. In this part, we used this implemented system since it is an effective and contrasted method for face detection, which is necessary to do before continuing with the rest of the face verification pipeline. Furthermore, using this detector, we can store the bounding boxes created by the algorithm which correspond to the coordinates where a face is detected, and we use this information in the tracking and identity assignment processes.

3.1.3. Change Shot Detection

The type of videos employed in this challenge are obtained from television programs, so these programs are usually composed of a huge variability in the content characteristics and constant changes of shot and scenes. Thus, to help the tracking and clustering step, we use a scene detection tool [2] which detects effectively these changes using the threshold-based detection mode. This detector finds areas where the difference between two subsequent frames exceeds a threshold value.

3.2. Embedding Extraction

Once the video processing step is done, we process the face images using the bounding boxes, apply mean and variance normalization, and resize to 160×160 pixels applying a central cropping. After that, the processed images are passed through a trained model to obtain embedding representations. In this system, as a face extractor, we have employed a pretrained convolutional neural network (CNN) with more than one hundred layers [14, 15]. This network was trained for a classification task on the CASIA-WebFace dataset [16], but the embeddings extracted from it have been proved previously in a verification task to check their discriminative ability with impressive results. For this reason, we decide to use these embeddings of 128 dimensions to extract the representations for enrollment and test files of this challenge.

3.3. Training Face Enrollment Models

Traditionally, in recognition tasks, a back-end is applied to compare enrollment and test embeddings and obtain the final verification scores. A widely used approach is cosine similarity where if an enroll identity has more than one enrollment embedding, these embeddings are averaged to compare with the test embedding. However, we demonstrated in [11] for the speaker verification task that a better solution to make this process consists of training an enrollment model for each enroll identity. Thus, in this work, we have applied this approach for face verification task where we have trained one model for each of the 161 enrollment identities. To train these models, we have used the embeddings of enrollment images, and video files from the development and test sets of the IberSPEECH-RTVE 2020 Challenge [10] as positive examples. While the enrollment files from the development and test sets of the IberSPEECH-RTVE 2018 Challenge [12] are used as negative examples.

Fig.2 shows the process to make this training where a learnable vector is obtained to represent each identity. This process consists of comparing positive or target examples with themselves (s_{tar}), and also with negative or non-target examples (s_{nontar}) using as training objective aDCF loss function [17]

[1] https://www.ffmpeg.org/

[2] https://www.pyscenedetect.readthedocs.io/en/latest/

which is an approximated function of a verification metric. To optimize aDCF loss, the scores used are obtained with cosine similarity.

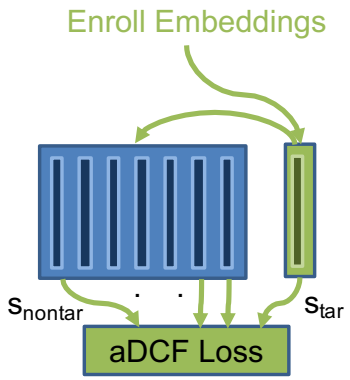


Figure 2: *Training face enrollment models using target and non-target embeddings.*

3.4. Clustering

As a source of complementary information, face embeddings from test videos are used to perform a spectral clustering technique [18] which tries to find strongly connected segments. This technique provides an initial cluster assignation to group the frames in the video sequence. In this work, we have employed this clustering combined with the use of the coordinates to improve the whole tracking process.

3.5. Tracking and Identity Assignment Scoring

Once all the above information is obtained, we have developed an algorithm to carry out the tracking and identity assignment process, which is depicted in Fig.1 and follows a similar philosophy to the one developed in [19]. In this algorithm, the tracking process has been developed by scene, so a shot change restarts the tracking. Therefore, while the scene is the same, the algorithm checks frame by frame the clustering information and the correspondence between the coordinates of the current frame and the previous frame to establish links which allow to make the tracking process. When a relation between both frames exists and has a high confidence term, the identity assignment of the previous frame is used to select the enrollment model and obtain the score. This score is compared with a confidence threshold to determine whether the identity assigned is correct or not. However, when there is no relation between the coordinates of the current frame and the previous frame or the confidence term is low, the frame embedding is compared with all the enrollment models to obtain a score and determine whether is a new identity to assign. Once the identity assignment is made over the current frame, the score is stored, the coordinates are updated, and the algorithm checks whether the scene changes.

Tracking is carried out with the previous steps, but the identity assignment process made is only an initial assignation. When a change of scene is detected, the system checks the identities and scores stored in the scene to remove inconsistent segment assignations. After that, the final segments with their identity assignations are written into the Rich Transcription Time Marked (RTTM) file. In addition, score confidence values are stored when a final identity assignment is made. If these values are greater than a more restrictive threshold which is set with the development set, we augment the enrollment models with the current face embedding. The whole process is repeated with all the detected scenes.

4. Speaker Subsystem

In this section, we present the speaker system which is based on similar blocks to the face system, such as audio processing, embedding extraction, clustering, and identity assignment scoring, but using different approaches in each one.

4.1. Audio Processing

4.1.1. Speech Activity Detection

Our approach for speech activity detection (SAD) is based on a deep learning solution which is an evolution derived from our previous experience with SAD systems in different domains [20]. We use a convolutional recurrent neural network (CRNN) consisting of 3 blocks of 2D convolutional (2 conv layer with 64 filters of size 3x3, batch normalisation and ReLU activation) followed by 3 BiLSTM layers. Then, the final speech score is obtained through a linear layer. As input features, 64 Mel filter banks and the frame energy are extracted from the raw audio and feed to the neural network. Cepstral Mean and Variance Normalization (CMVN) [21] normalization is applied.

4.1.2. Speaker Change Point Detection

The Speaker Change Point Detection block works in terms of Bayesian Information Criterion (BIC), according to its differential form (ΔBIC)[22]. We consider analysis windows of 6 seconds, modelling speakers with full-covariance Gaussian distributions. This block prioritizes those speech/non-speech boundaries given by SAD. As input features, the system considers 20 MFCC [23] features vectors, over a 25 ms hamming window every 10 ms. Features are then normalized according to CMVN to mitigate channel effects.

4.2. Embedding Extraction

Once the audio processing is done, each one of the obtained segments will be transformed into a compact representation also known as embedding. For this purpose, we have opted for an evolution of x-vectors [24] considering an extended version [25] of the TDNN architecture. Compared to the original, we have substituted the original mean and standard deviation pooling block by a multi-head self-attention block [26]. This self-attention block considers H different patterns, also known as heads, learnable from the own data. The output of the block consists of the concatenation of the estimated means and standard deviations. The neural network has been trained with data from VoxCeleb 1 [27] and 2 [28]. The resulting neural network provides embeddings of dimension 512. These embeddings will be later centered, dimensionsality reduced by means of LDA up to 200 and whitening and length-normalized [29].

4.3. Clustering

The obtained embeddings are modeled in a generative manner according to [30], where a tree-based PLDA clustering is proposed. This solution proposes a Maximum A Posteriori (MAP) estimation of the speaker labels Θ given the set of embeddings Φ. The model considers a Fully Bayesian PLDA [31] of dimension 100 to model $P(\Phi|\Theta)$, while the priors [32]. As we

did in [30], we interpret $P(\Phi|\Theta)$ as a tree structure by means of the product rule of probability. Hence, we opt for an optimization of the model according to a sequential manner making use of the M-algorithm [33] to find the best possible path along the tree. Moreover, prior to any clustering evaluation the PLDA model is adapted thanks to unsupervised adaptation approaches as described in [34].

4.4. Identity Assignment Scoring

The Identity Assignment (IA) block follows the schematic of a speaker verification task based on the standard embedding-PLDA paradigm. Hence, as preparation, each one of the enrollment recordings is converted into its corresponding embedding as well as the obtained segments from diarization. For the speaker verification task itself, enrollment models are build according to the correspondings audios while test models represent the clusters obtained during diarization. Each test model is made in terms of all segments assigned to the cluster. For simplicity reasons, we make use of the same embedding extractor and PLDA trained for diarization purposes.

After the scores are obtained, we normalized them using an adaptative s-norm. For each segment, we select cohorts similar to the test segment to compute the normalization values. For each trial, we selected 25% of the total segments in the cohort. The selection is based on the own PLDA scores. The final labels are built according to a threshold adjusted during calibration. This adjustment was obtained experimentally with the development set. Furthermore, as a design choice, we do not exclude the possibility of multiple clusters assigned to the same enrollment. This decision was made as to allow the correction of errors during diarization.

5. Results

In this section, the results for each subsystem are obtained using Diarization Error Rate (DER) as metric to evaluate. DER is usually the reference metric employed in diarization task, but in this case, DER is obtained slightly different than the original metric since it also takes into account the measurement of the identity assignment errors. Table 1 presents DER results obtained in the development and test set for face and speaker modalities. In addition to separate results, we show the mean result achieved with both systems. These results indicate a great mismatch between development and test results. We have analyzed which kind of video files composed both subsets and the length of these files, and we have found that development files are shorter than test files. Thus, we can see that the face and speaker subsystems obtain better performance in development files which are shorter videos, so the tracking process is easier to follow.

Table 1: *Experimental results on RTVE 2020 Multimodal Diariazation set, showing DER%. These results were obtained for the development and test sets in both modalities.*

Subset	Modality	DER%
DEV	FACE	51.66
	SPEAKER	47.90
	FACE+SPEAKER	49.78
TEST	FACE	61.79
	SPEAKER	72.63
	FACE+SPEAKER	67.21

To analyze better these results, Table 2 shows a decomposition of DER metric in the three terms of error:

- *Probability of misses (MISS)*: which indicates the segments where the target identity is presented but the system does not detect it.
- *Probability of false alarm (FA)*: which illustrates the number of errors due to the assignment of one enrollment identity to a segment without identity known.
- *Identity error (ID)*: which reflects the segments assigned to enrollment identities different from the target identity.

Focusing on face modality errors, in the case of development subset, we observe that the main cause of error is the probability of misses which indicates that a huge amount of segments from target identities have not been detected. Therefore, this effect can be motivated by the fact of using a threshold value too high. While in the test subset, misses and false alarm terms are similar. Especially relevant is the great increase of the false alarm errors since this fact illustrates the problems to discard segments of non-target faces when the number of enrollment identities is large. On the other hand, the distribution of errors produced in the speaker subsystem is quite different, because false alarms are much bigger than misses in both subsets of data. Note that it is also related to the threshold chosen. However, in this case, the threshold is lower, so the target segments are mostly detected, but as a result, a high number of enroll identities are assigned to segments of unknown identity.

Table 2: *Decomposition of DER% results in Miss (MISS), False Alarm (FA) and Identity (ID) Errors for the development and test sets in both modalities.*

Modality	Subset	MISS	FA	ID
FACE	DEV	37.5	6.5	7.7
	TEST	29.0	19.5	13.3
SPEAKER	DEV	14.0	29.6	4.3
	TEST	5.1	53.3	14.2

6. Conclusions

This paper presents the ViVoLAB submission to the IberSPEECH-RTVE 2020 Multimodal Diarization Challenge. In this work, we have developed two monomodal subsystems to address separately face and speaker diarization. Each system is based on state-of-the-art DNN approaches. We have demonstrated that there is still room for improvement in each of the systems because the results obtained are too high in both subsets and in both systems. Moreover, future work can be done on the fusion of both systems, which could improve the final results. The high DER values for misses and false alarms in the face and speaker subsystem, respectively, should be addressed by that fusion.

7. Acknowledgements

This work has been supported by the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the project TIN2017-85854-C4-1-R, by the Government of Aragon (Reference Group T36_20R) and co-financed with Feder 2014-2020 “Building Europe from Aragon”, and by Nuance Communications, Inc. The Titan V used for this research was donated by the NVIDIA Corporation.

8. References

[1] J. Poignant, H. Bredin, and C. Barras, "Multimodal person discovery in broadcast tv at mediaeval 2015," in *MediaEval 2015 working notes proceedings*. CEUR-WS. org, 2015.

[2] H. Bredin, C. Barras, and C. Guinaudeau, "Multimodal person discovery in broadcast TV at MediaEval 2016," in *MediaEval 2016 working notes proceedings*. CEUR-WS. org, 2016.

[3] O. Sadjadi, C. Greenberg, E. Singer, D. Reynolds, L. Mason, and J. Hernandez-Cordero, "The 2019 NIST Audio-Visual Speaker Recognition Evaluation," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 259–265.

[4] R. K. Das, R. Tao, J. Yang, W. Rao, C. Yu, and H. Li, "HLT-NUS Submission for NIST 2019 Multimedia Speaker Recognition Evaluation," *arXiv preprint arXiv:2010.03905*, 2020.

[5] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4930–4934.

[6] I. Viñals, P. Gimeno, A. Ortega, A. Miguel, and E. Lleida, "In-domain Adaptation Solutions for the RTVE 2018 Diarization Challenge," *Proc. IberSPEECH 2018*, pp. 220–223, 2018.

[7] E. Khoury, P. Gay, and J.-M. Odobez, "Fusing matching and biometric similarity measures for face diarization in video," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, 2013, pp. 97–104.

[8] N. Le, A. Heili, D. Wu, and J.-M. Odobez, "Efficient and Accurate Tracking for Face Diarization via Periodical Detection," in *International Conference on Pattern Recognition*, no. CONF. IEEE, 2016.

[9] A. Ortega, A. Miguel, E. Lleida, V. Bazán, C. Pérez, M. Gómez, and A. de Prada, "Albayzin evaluation: IberSPEECH-RTVE 2020 Speaker Diarization and Identity Assignment," 2020.

[10] E. Lleida, A. Ortega, A. Miguel, V. Bazán, C. Pérez, M. Gómez, and A. de Prada, "Albayzin evaluation: IberSPEECH-RTVE 2020 Multimodal Diarization and Scene Description Challenge," 2020.

[11] V. Mingote, A. Miguel, A. Ortega, and E. Lleida, "Training Speaker Enrollment Models by Network Optimization," *Proc. Interspeech 2020*, pp. 3810–3814, 2020.

[12] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. de Prada, "Albayzin 2018 evaluation: the iberspeech-RTVE challenge on speech technologies for spanish broadcast media," *Applied Sciences*, vol. 9, no. 24, p. 5412, 2019.

[13] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 815–823.

[15] D. Sandberg, "Face Recognition using Tensorflow," https://www.github.com/davidsandberg/facenet.

[16] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[17] V. Mingote, A. Miguel, D. Ribas, A. Ortega, and E. Lleida, "Optimization of False Acceptance/Rejection Rates and Decision Threshold for End-to-End Text-Dependent Speaker Verification Systems," *Proc. Interspeech 2019*, pp. 2903–2907, 2019.

[18] J. Shi and J. Malik, "Normalized cuts and image segmentation," Tech. Rep., 2000.

[19] E. Ramos-Muguerza, L. Docío-Fernández, and J. L. Alba-Castro, "The GTM-UVIGO System for Audiovisual Diarization," *Proc. IberSPEECH 2018*, pp. 204–207, 2018.

[20] I. Viñals, P. Gimeno, A. Ortega, A. Miguel, and E. Lleida, "Estimation of the Number of Speakers with Variational Bayesian PLDA in the DIHARD Diarization Challenge," in *Proc. Interspeech*, 2018, pp. 2803–2807.

[21] M. J. Alam, P. Ouellet, P. Kenny, and D. O'Shaughnessy, "Comparative evaluation of feature normalization techniques for speaker verification," in *International Conference on Nonlinear Speech Processing*. Springer, 2011, pp. 246–253.

[22] S. Chen, P. Gopalakrishnan *et al.*, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA broadcast news transcription and understanding workshop*, vol. 8. Virginia, USA, 1998, pp. 127–132.

[23] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[24] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 165–170.

[25] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, F. Richardson, S. Shon, F. Grondin *et al.*, "State-of-the-Art Speaker Recognition for Telephone and Video Speech: The JHU-MIT Submission for NIST SRE18." in *Interspeech*, 2019, pp. 1488–1492.

[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015*.

[27] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[28] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[29] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth annual conference of the international speech communication association*, 2011.

[30] I. Viñals, P. Gimeno, A. O. Giménez, A. Miguel, and E. Lleida, "ViVoLAB Speaker Diarization System for the DIHARD 2019 Challenge." in *INTERSPEECH*, 2019, pp. 988–992.

[31] J. Villalba and E. Lleida, "Unsupervised adaptation of plda by using variational bayes methods," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 744–748.

[32] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6301–6305.

[33] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[34] I. Viñals, A. Ortega, J. Villalba, A. Miguel, and E. Lleida, "Unsupervised adaptation of PLDA models for broadcast diarization," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, pp. 1–13, 2019.

The GTM-UVIGO System for Audiovisual Diarization 2020

Manuel Porta-Lorenzo, José Luis Alba-Castro, Laura Docío-Fernández

AtlanTTic Research Center, University of Vigo

mporta, jalba, ldocio@gts.uvigo.es

Abstract

This paper explains in detail the Audiovisual system deployed by the Multimedia Technologies Group (GTM) of the atlanTTic research center at the University of Vigo, for the Albayzin Multimodal Diarization Challenge (MDC) organized in the Iberspeech 2020 conference. This system is characterized by the use of state of the art face and speaker verification embeddings trained with publicly available Deep Neural Networks and fine-tuned for the persons of interest. Video and audio tracks are processed separately and are finally fused to make joint decisions on the speaker diarization result. Few modifications have been made over the GTM-UVIGO system presented in the very same conference in 2018, mainly regarding the video processing part.

Index Terms: speaker recognition, face recognition, deep neural networks, image processing, multimodal diarization.

1. Introduction

In recent years, the field of pattern recognition has witnessed a shift from the extraction of handmade features to machine-learned features using complex neural network models. Biometric verification is a clear example of an application scenario where Deep Neural Networks have produced a notable increase in performance, providing transformations of space where the face and voice of users are represented in clusters that are more compact and separable than in the original sample space. This representation makes the problem of diarization and verification in multimedia content more tractable than with previous approaches [1][2][3][4].

However, facial and speaker verification models are still not perfect and make many mistakes in verifying the identity of people in natural conditions. These situations are common when analyzing audiovisual content with frequent shot changes, camera movement, different types of scenarios, variability in the appearance of faces (pose, expression, illumination, blurring and small size), variability in the mix of voices, noise and background music. Also, the appearance of many other people who are not registered to be identified and are considered "intruders" to the system, causes many false identity assignments.

In this paper we explain the approach that the GTM research group has followed to tackle the person identification problem in audiovisual content. We have prepared a system that works separately on the video and audio tracks and makes a final fusion to fine tune the speaker diarization result. The rest of the paper is organized as follows. Section 2 explains the video processing part, including the segmentation of the video footage into different shots and the face detection, tracking, verification and post-processing at shot level. Section 3 explains the Speaker Diarization and Verification subsystem. Sections 4 and 5 present the experimental results Finally, Section 6 gives the computational cost information, and Section 7 presents the conclusions and details the on-going research lines.

2. Face diarization

Television programs such as news, debates, interviews, documentaries, etc., are characterized by frequent changes of shot and scene, the appearance of multiple people in foreground and also people and dynamic content in the background. On the other hand, other programs like TV-series can contain faces in very extreme appearances regarding pose, expression, illumination, make-up and size. Therefore, the final audiovisual content is very different from the typical scenarios where biometric identification is used, such as restricted access, video security or mobile scenarios.

The solution we have adopted for this competition in the video processing part is based on two fundamental ideas that apply to this type of content. On the one hand, we know that a change of shot implies, in general, a change in the person who appears in the scene, although it does not always happen and it does not happen in the same way regarding the speaker or the type of program. On the other hand, the people who appear in a shot remain in it as long as there is no movement of the camera or of the people themselves. This way, detection of shot changes gives an important clue for subsequent face processing. In the updated system for the 2020 challenge, television programs are much more diverse and the assumption that the camera is still most of the time in each shot does not hold anymore. Also, some of the programs have large digital screens at the background showing dynamic content that can mislead the former module of shot change detection. So, in the next section we explain the main changes applied to this module to cope with the high false detection rate produced by camera or background dynamics.

2.1. Detection of shot changes

This subsection explains a simple approach to detect shot changes designed to work in a ROC point with FP > FN. Shot changes will be used to restart face trackers because we cannot rely on tracking a face through shot changes, so losing a shot change could have a greater impact in the tracker than initializing the face tracker unnecessarily.

Detection of movement is also an important feature to have a more complete understanding of the footage, but we haven't included in this version of the system a specific movement detection block. Instead, we have used the false positive rate of the shot detection block as an indication of movement.

The steps to detect a change of shot are the following:

1. Reduce the size of the frame to save computational load,
2. Calculate the derivatives of the image to keep the edges of the scene,

10.21437/IberSPEECH.2021-17

3. Divide the frame into blocks and calculate the mean of edge pixels per block,
4. Subtract the mean of the same block in the previous frame,
5. Set a threshold for considering that a block difference represents a change (threshold set with the development video footage),
6. Count the number of block changes and set an upper and a lower threshold, defined for reliable changes of shot and shot continuity, respectively (also using the development set).
7. For the cases with values between both thresholds, detect all the faces in the frame and compare them with the detections of the previous frame in terms of position, size and distance between faces against a third threshold.

This "Canny-style" thresholding allows the system to track faces over smooth transitions and camera movements, potentially procuring longer tracks which are more suitable for the shot-based face processing. Nevertheless, these thresholds are quite dependent on the type of program and video realization, and so it is left for future improvements of the system the dynamic adaptation of those thresholds via the detection of different kinds of program. For this version a set of permissive values which minimize the number of false negatives for the whole ensemble of target videos was empirically obtained.

2.2. Face processing

The face processing subsystem comprises several sequential operations that are briefly explained through the Figure 1 and in the subsections below.

2.2.1. Face detection and geometric normalization

Face detection is a fundamental step in the sequential processing. We have used the detector based on Multi-Task Cascaded Convolutional neural Network [5], that jointly finds a Bounding Box for the face and five landmarking points useful to normalize the face. This face detector is quite robust to pose, expression and illumination changes. False negatives are typical in extreme poses with yaw angles beyond +/-60º and pitch angles beyond +/- 40º, that are not so uncommon in interview and debate contents, and quite typical in TV-series. This approach also brings a bit amount of false positives in areas where textured objects with skin colors appear, like hands, arms and other not human objects.

Once a face is detected (being true detection or not), its bounding box (BB) is saved with several parameters that will allow to do tracking and assign identities during the process. An overlapping function between the current BB, and the BBs of the previous frame allows linking the BBs belonging to the same person and processing full tracks when the shot has finished.

The detected face is passed to a geometric normalization that prepares the face to be plugged in in a standardized way to the face recognition block.

2.2.2. Training and fine-tuning a face recognizer

We have used the face recognizer based on dlib's implementation [6] of the Microsoft ResNet DNN [1]. This DNN finds an embedded space where faces with the same ID are grouped together and are far from faces with different ID.

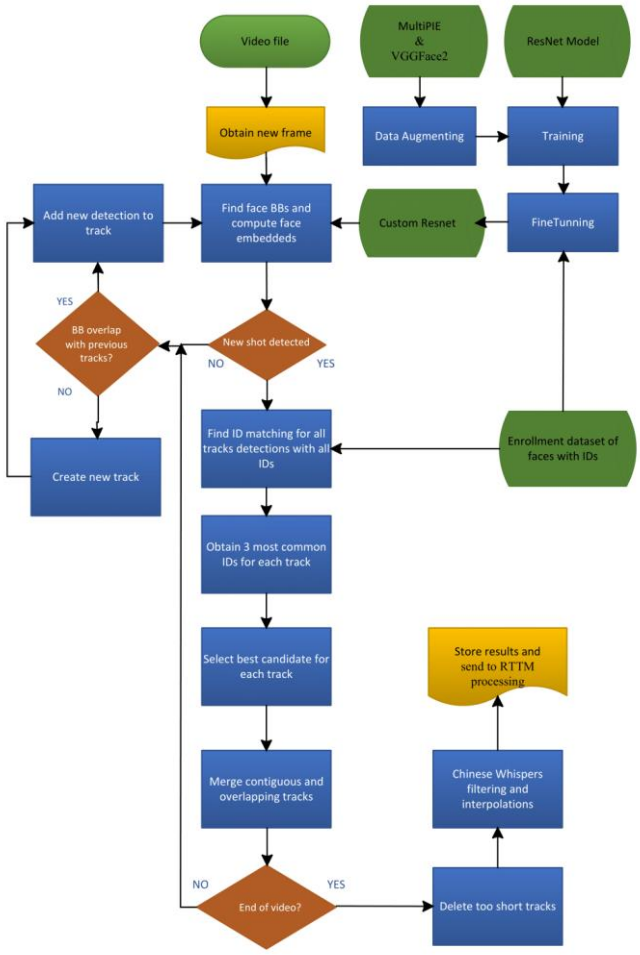


Figure 1. *Flow diagram of face processing.*

In order to improve the performance of the original dlib's network when poses are beyond +/-50º in yaw and +/- 20º in pitch, and face expressions are not neutral or smiling (the most typical in public datasets), we have retrained the network from scratch using the VGGFace2 [7] dataset and a subset of the MultiPIE face database [8] that contains extreme poses, not neutral expressions and non-uniform illumination. Re-scaling was also performed for augmenting small face samples. This network was then fine-tuned to pay more attention to the faces of the enrollment set. For the 2020 competition, 161 identities were provided, with 10 images per ID. Hand reviewing and normalization of incorrect aspect-ratio of many images was necessary to avoid the network learning aspect-ratios that would not appear in the training, development and test videos. Data augmentation was applied by horizontal flipping, XY-shifting, and downsampling-upsampling to provide blurred images. With this augmented image set a fine tuning of the previous network was programmed with the next characteristics:

- All weights of the DNN are allowed to learn through SGD with an initial learning rate of 1e-6. This decision worked better than freezing weights for some convolutional layers and allowing learning to the rest and the FC final layer.

- In order to avoid the catastrophic forgetting effect when fine tuning the Face recognizer, we have resorted to the rehearsal technique [9] and added face samples from the original training with VGGFace2 dataset. Not doing so, would make the model tending to memorize the enrollment IDs but failing more on unknown IDs, increasing False Alarm rate. Mini Batches of 100 identities and 10 samples per identity were randomly but evenly extracted from the augmented enrollment dataset and the VGGFace2 dataset, and trained with a cross-validation stop condition over the LFW (Labeled Faces in the Wild) dataset [14]. Fine tuning stops as soon as the FAR+FRR decreases a cumulative 1% over 5 steps (5000 face images). This method ensures that the face recognizer is tuned to the enrollment set but doesn't forget to tell apart also unknown IDs. It is important to note that the VGGFace2 dataset was reduced by taking out the LFW identities.

2.2.3. Shot-based face processing

When a face is detected in a shot its Bounding Box is compared with the last element of the current shot tracks and it is assigned to the closer one if it surpass a distance threshold or to a new one otherwise (or if there aren't any tracks on the current shot). As a shot change can make two faces belonging to different IDs to appear in the same position, a change of shot restarts the tracking process.

After a change of shot is detected, meaning that each stored track accumulates samples of the same person, a candidate ID is assigned to each track by comparing the embedded vectors of all its samples with all the embedded vectors of the enrollment set. For each sample, the closest enrollment ID is obtained and the three most frequent ones among all samples are taken into account. From these three candidates, the one which achieves the smallest individual distance between any pair of track and enrollment set samples is selected as the candidate ID and such distance is stored as the track reference value. Afterwards, in order to filter false positives and faces from persons who don't belong to the enrollment set, the reference value of the track is compared against a dynamic threshold which is defined to be less restrictive for the IDs with higher presence in the video so far. The rationale behind this method is to make the most of the tracking information, using the higher quality detections to select the candidate ID for all the samples in a track, and to ease the recognition of the persons having more presence in the video. Once a track has an assigned ID it is compared to the previously processed tracks of the same shot in terms of assigned ID, distance between bounding boxes and separation in time in order to detect fragments belonging to the same ID and join them by merging its detections. The processed tracks are stored and the program jumps to the next shot until the end of the video.

When the full video is processed the obtained tracks pass through different filters to enhance the performance of the system. First, the tracks which contain less than a fixed number of frames are deleted as they mostly correspond to detection errors. Following, the embeddings of each track are clustered using a Chinese Whispers algorithm [12] keeping only those belonging to the main cluster. Finally, the frame indexes and bounding boxes of each track's detections are interpolated to obtain continuous segments.

3. Speaker diarization

The used strategy for speaker diarization and verification is similar to those of the GTM-UVIGO system presented in the 2018 Challenge [13]. Specifically, it uses a DNN trained to discriminate between speakers, and which maps variable-length utterances or speech segments to fixed-dimensional embeddings that are also called x-vectors [2].

A pretrained deep neural network downloaded from http://kaldi-asr.org/models.html was used. The network was implemented using the nnet3 neural network library in the Kaldi Speech Recognition Toolkit [10] and trained on augmented VoxCeleb 1 [11] and VoxCeleb 2 data [15].

3.1. Speaker enrollment

The audio signal provided for each person in the enrollment set is used to obtain DNN speech-based embeddings. A sliding window of at least 10 seconds with a half a second hop is used. Then, these embeddings are clustered using the Chinese Whispers algorithm [12]. The threshold of the clustering algorithm is adjusted so that the clusters are pure and at least as many as the number of identities in the enrollment set. In this way an enrolled person can be represented by one or more clusters.

3.2. Off-line speaker diarization

First, the audio signal is divided into 3 second segments with a half a second hop. DNN short-term audio embeddings were extracted for each of these segment, clustered using the Chinese Whispers algorithm and their timestamps kept. From the clustering result we obtain an audio segmentation. Next, each of these segments, of arbitrary duration, are processed in order to extract one or more long-term audio embeddings using the same DNN. To do this, a sliding window of at least 10 seconds with a half a second hop is used. Then, these embeddings are clustered using again the Chinese Whispers algorithm, using a threshold that minimizes the diarization error.

3.3. On-line identity assignment

The clusters obtained in the previous step need to be assigned to the enrollment identities. Keeping the timestamps of each embedding in the clustering process, allows to design an online ID assignment approach. Time segments are defined as consecutive timestamps with embeddings associated to the same cluster. The ID assigned to a time segment is the enrollment ID of the best-matching enrollment cluster, as far as this distance is less than a threshold. This threshold is defined after observing the typical behaviour of the system in the development scenarios. A confidence value for that ID in that specific time segment is stored to be used jointly with the face-based confidence value in the fusion process.

3.4. Fusion

The SPEAKER modality in the 2020 contest with 161 IDs of enrollment, produced too many false assignments of time segments. To correct potentially wrong speech-based ID assignments, a multimodal fusion approach that uses the assignments made by both modalities separately was implemented. Given a time segment that has been assigned a speaker identity ID1, three rules are applied depending on the FACE modality content:

1. If no faces were assigned to any enrollment ID in the same time segment, and also, the identity ID1 is not found anywhere in the video using the face modality, the speaker ID1 is removed from the speaker output file. That is, it is very unlikely that when someone speaks, no face will appear in the video in that time interval and, furthermore, that identity will not appear in the whole video.
2. If a high-confidence single face identity ID2 has been detected in more than 60% of the video frames in that speaker ID1 time segment, and the identity ID1 is not found anywhere in the video using the face modality, the speaker ID1 is changed to the face identity ID2. That is, it is very unlikely that someone will speak on the video and never appear his face.
3. If several face identities (including ID1) were assigned with high-confidence in the same time segment that the speaker identity ID1, and a single face identity ID2 has been detected in enough frames (above the 60%), the assigned ID1 is changed to the face identity ID2. This rule doesn't apply if ID1 and ID2 have different gender (as given by the enrollment name). This makes the face modality more reliable than the speaker one.

4. Results on Development videos

The primary evaluation metric to rank systems is the average of the face and speaker diarization errors (Averaged DER). Performance metrics over one of the Development videos provided by the organizers of the competition are presented in Table 1. It is worth noting that we consider this video as a testing one, that is, the enrollment identities to search were those of the Test dataset.

Table 1: *Results on one of the Development videos*

Modality	DER	MISSED	FALARM	ERROR
Face (F)	15.77	11.9	3.8	0
Speaker (S)	32.01	9.2	20.3	2.5
Speaker Fusion (SF)	24.81	9.6	15.2	0
Averaged F & SF	20.29	10.75	9.5	0

5. Results on Tests videos

The results over the Test videos provided by the organizers of the competition are presented in Figure 2. In this graph, the speaker DER refers to the modified output after fusion with Face output. X axis shows the acronym for the program and Y-axis the DER (face - BLUE, speaker fusion - RED and average of both - ORANGE). The system meta-parameters have been adapted to the details of the AT ("Aquí la Tierra") program used for developing, so it seems clear the dependency of the system to the type of program, especially in the SPEAKER modality, but also in FACE. It is also noticeable the bad performance of SPEAKER compared to FACE in the SFT ("Si Fueras Tu") program.

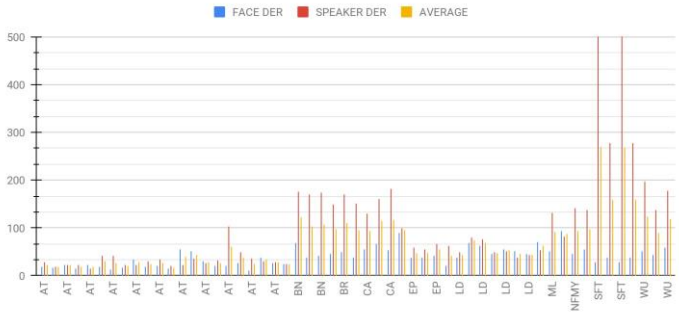


Figure 2. *DER results on Test videos*

6. Computational cost

The computational cost of the proposed audiovisual diarization system was measured in terms of the real-time factor (RT). This measure represents the amount of time needed to process one second of audiovisual content: xRT = P/I, where I is the duration of the processed video and P is the time required for processing it. An example video (AT-20181111.mp4) was processed to compute the RT, thus taking into account many different audiovisual situations. The duration of this video is I = 1743.72 s, and the time needed to process it was P = 7229.5 s, leading to RT = 4.146. This computation time was obtained by running this experiment on an Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz with 256 GB RAM. Even though the process is running more than 4 times slower than real-time, the code is not optimized at all (it is completely implemented in Python and the machine is not fully exploited.

7. Conclusions and futures work

We have presented the GTM-UVIGO System deployed for the Albayzin Multimodal Diarization Competition at Iberspeech 2020. The system uses state of the art DNN algorithms for face detection and verification and also for speaker diarization and verification including fine tuning of the face recognition model using the persons of interest. In order to minimize tracking errors, a shot detection algorithm resets the face trackers and makes use of a Canny-style double threshold to optimize the change decision. A novel shot-based faces tracking is also proposed, which makes the most of the temporal information, retrieves low quality faces, filters false positives using a dynamic threshold and provides continuity to the output detections. The application scenario is studied to implement ad-hoc post-processing strategies to fine-tune the ID assignments made by the video and audio parts. This framework leaves a lot of room for improvement in each of the fundamental processing stages and also in the ad-hoc rules for post shot fine-tuning.

8. Acknowledgements

This work has received financial support from the Xunta de Galicia (Agrupación Estratéxica Consolidada de Galicia accreditation 2016-2019; AtlanTTic and ED431B 2018/60 grants) and the European Union (European Regional Development Fund ERDF).

9. References

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.

[2] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” in INTERSPEECH 2017 – 97th Annual Conference of the International Speech Communication Association, Proceedings, pp. 999–1003, 2017.

[3] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, “End-to-end text-dependent speaker verification,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 5115–5119.

[4] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey and A. McCree, "Speaker diarization using deep neural network embeddings," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 4930-4934.

[5] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, “Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks,” IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, Oct. 2016.

[6] D. E. King. “Dlib-ml: A Machine Learning Toolkit, Journal of Machine Learning Research,” 10:1755-1758, 2009.

[7] Cao, Qiong, Shen, Li, Xie, Weidi, Parkhi, Omkar and Zisserman, Andrew. (2018). VGGFace2: A Dataset for Recognising Faces across Pose and Age. 67-74. 10.1109/FG.2018.00020.

[8] http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html

[9] Robins, Anthony. "Catastrophic Forgetting, Rehearsal and Pseudorehearsal". Connection Science. 7 (2): 123–146.1995

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., “The Kaldi speech recognition toolkit,” in IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 1-42011.

[11] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: a large scale speaker identification dataset,” INTERSPEECH 2017 – 97th Annual Conference of the International Speech Communication Association, 2017.

[12] Chris Biemann, “Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems,” in First Workshop on Graph Based Methods for Natural Language Processing (TextGraphs-1), pp. 73-80, 2006.

[13] Eduardo Ramos-Muguerza, L. Docío-Fernández and J. L. Alba-Castro. “The GTM-UVIGO System for Audiovisual Diarization”, In Proceedings of the IberSPEECH, 2018; pp. 204–207.

[14] G.B. Huang, M. Ramesh, T. Berg and E. Learned-Miller. “Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments”, University of Massachusetts, Amherst, Technical Report 07-49, 2007.

[15] J. S. Chung, A. Nagrani, A. Zisserman. “VoxCeleb2: Deep Speaker Recognition”, INTERSPEECH, 2018.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

The Biometric Vox System for the Albayzin-RTVE 2020 Speaker Diarization and Identity Assignment Challenge

Roberto Font[1], *Teresa Grau*[1]

[1]Biometric Vox S.L.

roberto.font@biometricvox.com, teresa.grau@biometricvox.com

Abstract

This paper describes the systems developed by Biometric Vox for the Albayzin Speaker Diarization Challenge organized as part of the Iberspeech 2020 conference. The two systems (primary and contrastive) we developed for the challenge are based on Deep Neural Network x–vector embeddings and a Probabilistic Linear Discriminant Analysis (PLDA) backend. The resulting x-vectors are grouped using Agglomerative Hierarchical Clustering (AHC) in order to obtain the diarization labels. Systems differ in the resegmentation stage. Our primary system achieves 14.96% DER on the test set of the RTVE2018 database and 21.35% on the 2020 evaluation set.

Index Terms: speaker diarization, speaker embeddings, x-vectors, speaker identification

1. Introduction

Speaker diarization is the task of segmenting a speech audio, marking speaker change points and categorizing those segments according to the speaker identity; in other words, speaker diarization answers the question of who speaks when.

The aim of this paper is to provide a complete description of the systems developed by Biometric Vox for the Albayzin-RTVE 2020 Speaker Diarization and Identity Assignment Challenge. The task for this challenge is to group together all speech segments belonging to the same speaker and, in case the speaker is known to the system, assign the corresponding identity. The material for the challenge consists in TV shows that belong to diverse genres and broadcast by the public Spanish Television (RTVE), covering a wide range of realistic and challenging conditions: spontaneous and read speech, different accents, background noise, songs, laughs, yells, quick short-turn conversations, etc.

Our two systems are based on the DNN x–vector paradigm [1] and share the same basic building blocks, which are described in more detail in Section 3:

- Acoustic feature extraction (23 MFCC features) and energy-based Voice Activity Detection (VAD).
- X–vector embedding extraction.
- Embedding post-processing: length-normalization, centering, whitening, Linear Discriminant Analysis (LDA).
- Probabilistic Linear Discriminant Analysis (PLDA) scoring.
- Agglomerative Hierarchical Clustering.

The primary system performs two diarization stages: the segments resulting from the first one are fed to a music/speech/noise classifier and music segments are removed. Once music segments are excluded, a second diarization stage is performed to obtain the final result. This system achieves a Diarization Error Rate (DER) of 21.35% on the 2020 evaluation set.

The contrastive system focuses on refining the speaker change boundaries, which we have found that can improve performance when there are long-turn conversations with no frequent speaker changes. This system obtains 31.57% DER on the 2020 evaluation set.

The rest of the paper is organized as follows: in Section 2 the RTVE database, the data used to train the embedding extractor and the data used as development and evaluation sets are described. Section 3 presents the basic components of our two systems and Section 4 describes the diarization process and characteristics for those systems. Section 5 covers our submitted system for the identity assignment task. Results are shown in Section 6, and, finally, we summarize our results and conclusions in Section 7.

2. Data Resources

2.1. RTVE database

The RTVE2018 database [1] has a total of 569 hours and 22 minutes of audio extracted from 17 different TV shows broadcast by RTVE (Radio Televisión Española) from 2015 to 2018. Most shows are related to news, debates, social gatherings and documentaries. The database is divided into 4 subsets: a training set, 2 development sets, *dev1* and *dev2*, and a test set, as summarized in Table 1. Around 37 hours, divided among the *dev2* set and a subset of the test set, have diarization and speech segmentation labels available. Additionally, the *dev2* set counts with several audio files for a limited set of speakers to allow for the creation of speaker models.

Table 1: *Composition of the different RTVE subsets.*

Subset	# hours	# different shows	# speaker models
RTVE2018			
train	500h	16	-
dev1	52:31:51	5	-
dev2	15:09:25	2	34
test	39:07:15	8	39
RTVE2020			
dev	03:55:31	2	18
test	67:23:29	15	161

The RTVE2020 database [2] is also a collection of TV shows from the public Spanish Television (RTVE) that were broad-

[1]http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf
[2]http://catedrartve.unizar.es/reto2020/RTVE2020DB.pdf

10.21437/IberSPEECH.2021-18

cast from 2018 to 2019. The database is composed of 70 hours and 18 minutes of audio belonging to 15 different shows (dev's shows are also in the test set). Only two shows are in common with the 2018 dataset: *Comando Actualidad* and *Millenium*. The database is divided into 2 subsets: *dev* and *test*. The latter one is the challenge evaluation set. Shows' genre for this set is related to entertainment: series, sketches, reports, entertainment shows, etc.

TV show's genre and format are quite different between the two sets; each set covers a different range of conditions. For the RTVE2018 database: spontaneous and read speech, long-turn conversations, different accents and background noise. And for RTVE2020: songs, laughs, yells, quick short-turn conversations and colloquial vocabulary.

2.2. Training Data

Our training material consists of the following datasets:

- NIST SRE 04-10
- MIXER6 as prepared by the Kaldi sre16 recipe.
- Switchboard phase1-3 and cellular1-2.
- Voxceleb 1 and 2. [2] [3]

We augment training data by generating four additional perturbed versions of each file using the Musan [4] corpus by adding:

- Reverberation
- Musan noise
- Musan music
- Musan speech

For the embedding extractor training, the training set consists of NIST SRE 04-10, MIXER6, Switchboard, VoxcelebCat and all augmented data from these datasets. VoxcelebCat is the result of concatenating all excerpts from the same video into one longer file and combining Voxceleb 1 train, Voxceleb 2 dev and Voxceleb 2 test. The samples from NIST SRE 04-10, MIXER6 and Switchboard were upsampled to 16kHz.

The total number of utterances is $1,109,458$ from a total of $13,682$ speakers.

2.3. Development and Evaluation Data

In order to evaluate the speaker diarization performance of our systems, we used the datasets provided by the organizers for this challenge:

- RTVE2018DB dev2, with 12 episodes from 2 shows, which are manually transcribed and with speaker time references for all episodes and identity references for one episode. Includes an enrollment set with audio files provided for 34 speakers.
- RTVE2018DB test, with 61 episodes from 9 shows. A subset of 40 of these episodes have speaker time references for speaker diarization available. The enrollment set has 39 speakers.
- RTVE2020DB dev, with 9 episodes from 2 shows with speaker time references and identity references for some of the speakers in each episode. Also includes an enrollment set with audio files provided for 18 speakers.

The 2020 evaluation set is composed of 87 episodes from 15 shows. Of those, 54 episodes from 10 shows are used for the speaker diarization challenge. Additionally, an enrollment set for 161 speakers is provided.

3. System Components

The diarization system, based on x-vector neural embeddings [1] and a PLDA backend, was implemented using the Kaldi [5] toolkit. The rest of this section provides a more detailed description of the main building blocks.

3.1. Feature Extraction and Voice Activity Detection

The input of the embedding extractors are 23-dimensional Mel Frequency Cepstral Coefficients (MFCCs) which are extracted from 25 ms windows with 15 ms overlap. Features are normalized using cepstral mean and variance normalization over a sliding-window of 300 frames.

Initial segmentation and silence removal is made using Kaldi standard energy-based VAD.

3.2. Embedding Extractor

To train the embedding extractor we used the baseline TDNN x-vector architecture as in the Kaldi SRE 16 recipe (Table 2).

Table 2: *Baseline x-vector architecture.*

Layer type	**Layer context**	**Size**
TDNN-ReLU-batchnorm	t-2:t+2	512
TDNN-ReLU-batchnorm	t-2, t, t+2	512
TDNN-ReLU-batchnorm	t-3, t, t+3	512
ReLU-batchnorm	t	512
ReLU-batchnorm	t	1500
Stats Pooling (mean+stddev)	T	2x1500
ReLU-batchnorm		512
ReLU-batchnorm		512
Softmax		# speakers

The model was trained for 3 epochs, with batch size of 64, on an Nvidia GeForce GTX 2080.

3.3. Back-end

All systems use a back-end that follows a classical LDA-PLDA scoring scheme:

- Embeddings are projected to unit length, centered and whitened.
- Linear discriminant analysis (LDA) is used to project the embeddings to a lower dimension. (From 512 to 150 in our case.)
- The segments are scored using PLDA.

Both LDA and PLDA are trained on VoxcelebCat. No score normalization or calibration was applied.

4. Diarization Systems

In this section, we provide a description of our submitted diarization systems. As discussed above, both systems share the same basic building blocks. However, the primary system performs two diarization stages to try and remove all music segments, which is a challenging feature of the RTVE dataset, while the contrastive system performs a resegmentation around the speaker change points to refine the segment boundaries.

4.1. Primary system

The speaker diarization process starts with feature extraction: 23-dimensional MFCC features are extracted, and Cepstral Mean Variance Normalization (CMVN) sliding window is applied before performing a VAD segmentation to remove silence portions.

DNN x-vectors embeddings are extracted for the resulting segments. A sub-segmentation with a sliding window of 3 seconds and a second and a half hop is used. The embeddings are clustered using Agglomerative Hierarchical Clustering (AHC) with single linkage.

The segments resulting from this first diarization stage are passed through a music/speech/noise classifier and music segments are removed. Finally, a second diarization stage is performed to obtain the final diarization result. Figure 1 shows the flowchart of our primary diarization system.

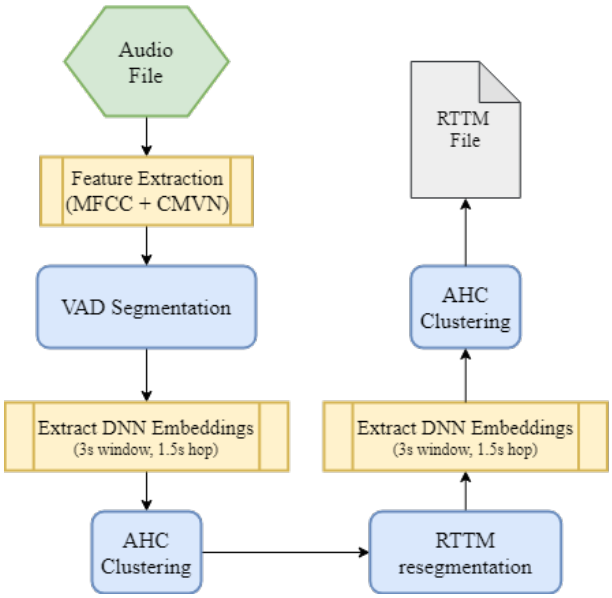


Figure 1: *Primary system's flowchart*

4.2. Contrastive system

The contrastive system uses the same feature extraction and VAD segmentation than the primary system, but no subsegmentation is used for embedding extraction. Embeddings are extracted for the whole segments resulting from the VAD segmentation. This embeddings are clustered using Agglomerative Hierarchical Clustering with single linkage. Then, a re-segmentation algorithm refines the speaker transition boundaries: DNN x-vectors embeddings with a sliding window of 3 seconds and a second and a half hop are extracted only for the segments involved in a speaker transition. The embeddings are then clustered until the number of clusters is two. Figure 2 illustrates the process to refine the speaker boundaries.

5. Identity Assignment

If there is prior knowledge about the identity of the people involved in an audio, it can be used to assign a name to each diarization label output by the speaker diarization system.

The steps for the identity assignment process in our submitted system are the following:

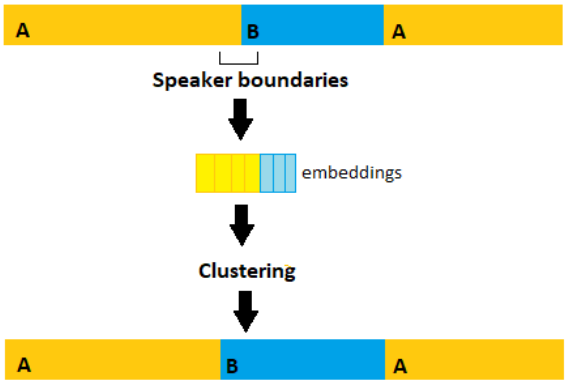


Figure 2: *Speaker boundaries refinement*

1. Creating speaker models. To that end, DNN x-vector embeddings are extracted from the audio files provided for each person in the enrollment set and averaged so that each speaker model is the average of the embedding for all enrollment files available for that speaker.
2. Performing speaker diarization with our primary system.
3. Extracting an x-vector embedding for each final segment obtained as a result of the diarization process.
4. Comparing each segment's x-vector with the speaker models by computing log-likelihoods ratios using the same back-end described in Section 3.3.
5. Assigning the identity of the best-matching speaker model to the segment only if it exceeds a threshold that was tuned on the development set.

6. Results

Table 3 shows the results for our two systems on the development and test portions of the RTVE2018DB.

Table 3: *DER (%) on the RTVE2018DB. MISS column is for the missed speaker time, FA for false alarm speaker time and SPK for speaker error time.*

System	MISS	FA	SPK	DER
Dev2				
Primary	2.1	1.4	4.1	7.68
Contrastive	1.9	1.7	7.0	10.59
Test				
Primary	2.7	3.5	8.7	14.96
Contrastive	0.6	3.9	14.6	19.11

The evaluation results provided by the organizers for the 2020 Speaker Diarization Challenge's evaluation set are shown in Table 4. Table 5 presents the 2020 evaluation results by TV show.

The best results are obtained for the shows *Millennium* (ML) and *Los desayunos de TVE* (LD), which revolve around news content and long-turn conversation almost without interruptions between speakers. For the entertainment domain, on the other hand, results suggest that improving the speech detection and music/noise removal stage could reduce the miss

speech and false alarm. This could also help to obtain a better segmentation and reduce the speaker error time.

For the identity assignment task, two additional metrics are used to evaluate the system performance: Assignment Error Rate (AER) and Average Speaker Error (ASE). For our system, the AER on the test data in the RTVE2018 dataset was 37.39% and for the ASE, 39.17%. The DER for the identity assignment evaluation set provided by the organizers is shown in Table 6. These results by show are consistent with those obtained for the speaker diarization task.

The processing time for the subset dev2 was 6 hours and 31 minutes on an Intel(R) Core(TM) i7-5820K CPU @ 3.30GHz with 6 cores.

Table 4: *DER (%) RTVE2020 Evaluation results. MISS column is for the missed speaker time, FA for false alarm speaker time and SPK for speaker error time.*

System	MISS	FA	SPK	DER
Primary	4.0	4.8	12.5	21.35
Contrastive	2.7	10.3	18.5	31.57

Table 5: *DER (%) RTVE2020 Evaluation results by TV show on our primary and contrastive systems for the Speaker Diarization task. TV shows: Aquí la Tierra (AT), Boca Norte (BN), Bajo la Red (BR), Comando Actualidad (CA), Ese Programa del que Usted me Habla (EP), Los desayunos de TVE (LD), Millennium (ML), Never Films Mira Ya (NFMY), Si Fueras Tu (SFT) and Wake-Up (WU)*

TV Show	Primary	Contrastive
AT	18.70	34.51
BN	100.70	137.54
BR	70.08	109.17
CA	35.55	55.84
EP	24.26	39.38
LD	13.60	12.56
ML	10.93	12.99
NFMY	64.63	146.97
WU	78.83	133.52

Table 6: *DER (%) RTVE2020 Evaluation results by TV show for the Identity Assigment task.*

TV Show	Primary
AT	97.33
BN	153.52
BR	123.18
CA	100.67
EP	64.00
LD	49.32
ML	80.28
NFMY	126.38
WU	112.03
Global	65.09

7. Conclusions

We have presented our diarization systems submitted to the Albayzin Speaker Diarization and Identity Assignment Challenge and reported the results on the development and evaluation sets. This challenge focuses on grouping together all speech segments belonging to the same speaker on TV shows, and assigning the identity in case the speaker is known to the system, a highly challenging task especially in the detection of the number of speakers and dealing with a wide range of realistic conditions such as background noise/music and overlapped speakers.

We submitted two systems for the speaker diarization task and one for the identity assignment subtask. The primary system focuses on performing a better music removal stage, which helps on the clustering stage to obtain better results. The contrastive system focuses on refining the boundaries between speakers. Unsurprisingly, this latter process, geared towards conversations with long turns, a condition that is not present in most of the shows under consideration, obtained a higher DER than our primary, more general-purpose, system.

8. Acknowledgements

The authors would like to thank the organizers of the Albayzin Challenge.

9. References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," 04 2018, pp. 5329–5333.

[2] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[3] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[4] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," 2015.

[5] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

The CLIR-CLSP System for the IberSPEECH-RTVE 2020 Speaker Diarization and Identity Assignment Challenge

Carlos Rodrigo Castillo-Sanchez[1], *Leibny Paola Garcia-Perera*[2]

[1]Computational Learning and Imaging Research, Universidad Autónoma de Yucatán, Mexico
[2]Center for Language and Speech Processing, Johns Hopkins University, USA

{carloscastillomvc,leibny}@gmail.com

Abstract

This paper describes the Speaker Diarization system jointly developed by the Computational Learning and Imaging Research (CLIR) laboratory of the Universidad Autónoma de Yucatán and the Center for Language and Speech Processing (CLSP) of the Johns Hopkins University for the Albayzin Speaker Diarization and Identity Assignment Challenge organized in the IberSPEECH 2020 conference. The Speaker Diarization system follows an x-vector-PLDA-VBx pipeline built with the Kaldi toolkit. It uses a Time Delay Neural Network (TDNN)-based Speech Activity Detector (SAD), with x-vectors as acoustic features, clustered with Agglomerative Hierarchical Clustering (AHC) as initialization for variational Bayes clustering. The system was only evaluated in the Speaker Diarization condition.
Index Terms: speaker diarization, time delay neural network, x-vector, vbx

1. Introduction

IberSPEECH's Albaizyn evaluation challenges cover a wide range of speech processing technologies that include speech-to-text transcription, search on speech, and speaker diarization, with the latter being the subject of this paper. Speaker diarization is the process of grouping the same speaker's utterances in an audio recording under the same label with no prior knowledge of the number nor identity of the intervening speakers. It is an essential preprocessing step for many speech applications, such as Automatic Speech Recognition (ASR), spoken document retrieval, or audio indexing [1]. Therefore, the improvement of speaker diarization technologies is crucial to perform adequately in real-world conditions. The IberSPEECH-RTVE Speaker Diarization and Identity Assignment Challenge calls for robust speaker diarization systems for real TV broadcast shows from a range of topics on the Spanish public network [2, 3].

In the previous Albaizyn evaluation, five teams submitted systems for the open-set speaker diarization condition. The ODESSA team [4] explored three different segment representation embeddings: Binary key, Triplet-loss, and x-vectors; trained with the challenge's data [5], NIST SRE, and VoxCeleb1, respectively. Their primary submission consisted of fusing at similarity matrix level three systems, one for each embedding type and clustering with AHC. This was possible as they shared the same 1-second segmentation.

The JHU team [6] also leveraged score fusion at similarity matrix level. They addressed four types of embeddings extractors: x-vector-basic, x-vector-factored, i-vector-basic, and bottleneck features (BNF) i-vector. The first extractor was trained on VoxCeleb1 and 2 with augmentations; the second one with SRE12-micphn, MX6-micphn, VoxCeleb and, SITW-dev-core; the third one with VoxCeleb1 and 2 with no augmentations; and the last one with the same data as x-vector-factored. Their pipeline used a TDNN-based SAD and Probabilistic Linear Discriminant Analysis (PLDA) trained with Albayzin2016 data. The four embeddings' similarity scores were fused on equal weights and clustered with AHC.

Our system follows the conventional diarization pipeline [7, 8, 9], described as follows: (1) Segmentation: in this step, the non-speech portions of the recording are removed, and the remaining speech regions are further cut into short segments. The system leverages a pre-trained, publicly available SAD[1] based on a TDNN with stats pooling. (2) Embedding extraction: in this step, the system extracts speaker-discriminative embedding for each segment; the submitted system uses x-vectors. (3) Clustering: after an embedding is extracted from each segment, the segments are grouped into different clusters; our system was tested with three different PLDAs based on in- and out-domain data, with AHC as initialization for variational Bayes clustering.

The paper is organized as it follows: in Section 2 the used databases are described. Section 3 further describes the system characteristics, and Section 4.1 presents the computational resources used.

2. Datasets

Four datasets were used to develop our speaker diarization system:

- VoxCeleb1: a large-scale speaker identification dataset with 1,251 speakers and over 100,000 utterances, collected "in the wild" [10].
- VoxCeleb2: a speaker recognition dataset that contains over a million utterances from over 6,000 speakers under noisy and unconstrained conditions [11].
- DIHARD II: focused on "hard" speaker diarization, contains 5-10 minute English utterances selected from 11 conversational domains, each including approximately 2 hours of audio [12].
- The Corporación Radiotelevisión Española (RTVE) speaker diarization database [3]: consists of around 70 hours of audio documents annotated in terms of speaker turns. A 41% of the database is used for evaluation purposes with no a priori information provided about the number of speakers. The remaining 59% consists of a collection of 8 different TV shows by the Spanish TV station provided in 3 partitions that can be used for system training and development.

[1]http://kaldi-asr.org/models/m12

10.21437/IberSPEECH.2021-19

The x-vector extractor model for the speaker diarization condition was trained using VoxCeleb1+2 augmented with DIHARD II data. The three tested PLDA models were developed as follows: The first PLDA model uses a mixture of the DIHARD II dev and RTVE 2018 datasets, augmented with MUSAN [13] noises and reverberation; the second one was trained with VoxCeleb1+2 data, with the last PLDA being the weighted interpolation of the previous two.

3. System overview

This section describes the components that comprise the developed speaker diarization system.

3.1. Speech activity detection

In order to extract speech segments, our system uses a pretrained TDNN-based SAD model[1]. The model was developed with the CHiME-6 training data [14]; such data was recorded in real-life conditions containing large amounts of background noise and overlapping speech. The SAD neural network architecture employs high-resolution Mel-Frequency Cepstral Coefficients (MFCC) as input, extracted for a 25ms window with a 10ms frame rate; with average log of energy, 40 mel-frequency bins, and a low cutoff frequency mel bins of 40. The network consists of 5 TDNN layers and two layers of statistics pooling [15]; trained with cross-entropy objective function to produce speech and non-speech labels. The speech labels include clean voice and voice with noises. Music, noise, and silence are categorized as non-speech. SAD labels are obtained by Viterbi decoding using an HMM with minimum duration constraints of 0.3 s for speech and 0.1 s for silence. We also tried energy-based SAD, but it was discarded as it performed worse overall.

3.2. Embeddings

We explored two types of embeddings. The first one, i-vectors; following the default Kaldi recipe for DIHARD, we trained a T-matrix with RTVE 2018-only data; afterward, we extracted i-vectors from the RTVE 2020 dataset and obtained baseline results. These i-vectors were of dimension 128.

The second type of embeddings we tested was x-vectors [16, 17]. We explored two methods to compute them; first, we followed the default Kaldi recipe for DIHARD, using VoxCeleb1+2 and RTVE 2018 with additional augmentation using MUSAN noises[2] as described in [13], to train the TDNN-based embedding extractor. This method passes each MFCC through a sequence of TDNN layers. A pooling layer computes the mean and standard deviation of the TDNN output over time, accounting for the utterance level process, with this internal representation (the x-vector) projected to a lower dimension. The DNN output is the training speakers' posterior probabilities, with the objective function being cross-entropy.

We used a TDNN-based extractor that uses 40-dimensional filterbanks with a 25ms window and 15ms frame shift as acoustic features for the second approach. These features are used for the embedding extraction as in [7]. The x-vector extractor model was trained using a TDNN with a 1.5s window with a frame shift of 0.25s; its architecture consists of four TDNN-ReLU layers, each of them followed by a dense-ReLU; afterward, two dense-ReLU layers are incorporated before a pooling layer; with a final dense-ReLU incorporated from which 512-dimensional embeddings are extracted. Then, a dense-softmax provides the output layer for this TDNN architecture.

Table 1: *x-vector extractor architecture [18].*

Layer	**Layer context**
frame 1	[t - 2, t - 1, t, t + 1, t + 2]
frame 2	[t]
frame 3	[t - 4, t - 2, t, t + 2, t + 4]
frame 4	[t]
frame 5	[t - 3, t, t + 3]
frame 6	[t]
frame 7	[t - 4, t, t + 4]
frame 8	[t]
frame 9	[t]
stats pooling (frame7, frame9)	[0, T]
segment1	[0, T]
softmax	[0, T]

3.3. PLDA scoring

As mentioned in Section 2 we tested our system with three different PLDA models; the first one was trained on a mixture of both DIHARD II dev and RTVE 2018 data augmented with MUSAN [13] noises and reverberation. The second PLDA used VoxCeleb1+2 out-of-domain data for training. For our third PLDA model, we followed [18]; this method aims to compute a robust PLDA based on the mixture of in-domain and out-of-domain PLDAs. This PLDA results from a weighted interpolation of the VoxCeleb1+2 out-of-domain data PLDA and the in-domain RTVE 2018+DIHARD II dev mixture PLDA. Both PLDAs were centered, whitened, and length normalized using the RTVE 2018+DIHARD II dev mixture data. Finally, the x-vectors were projected from 512 dim to 220 using Linear Discriminant Analysis (LDA).

3.4. Clustering

Using the similarity scores from one of the PLDAs, an Agglomerative Hierarchical Clustering (AHC) algorithm creates a set of clusters with an overestimation in the number of speakers. The VBx [19] uses the AHC initialization to make a further refinement of the clusters. VBx eliminates redundant speakers across the recording; it can be tuned by modifying the regulation coefficient (aggressiveness of eliminating redundant speakers), the acoustic scaling factor, and the loop-probability (staying in the same state when getting the next observation). The values used for our system are 0.4, 11, 0.80, respectively.

4. Experiments

This section describes some of the experiments that took place during our system's development process. We evaluated our systems using two metrics; the first one was Diarization Error Rate (DER), the most common metric for speaker diarization. DER comprises four types of errors: speaker error, false alarm speech, missed speech, and overlap speaker. Our DER followed the IberSPEECH-RTVE's evaluation plan characteristics, having a forgiveness collar of ±0.25 s before and after each reference boundary; and consecutive segments of the same speaker with a silence of less than 2 s come together as a single segment. The second metric that gave us an idea of the systems' performance was speaker number error; it allowed us to observe how each system estimated the number of speakers for each record-

[2]http://www.openslr.org/resources/17

Table 2: *DER (%), speaker error (SE) (%), missed speaker (MS) (%), false alarm (FA) (%) and speaker number error (%) comparison of different setups for the RTVE dev dataset (**post-submission** results are in bold letters). The speaker number error is the mean absolute error of the inferred number of speakers per recording.*

System	alpha, fa, fb, p	DER	SE	MS	FA	Speaker # error
i-vectors + DIHARD/RTVE PLDA + AHC	-	85.55	29.05	48.00	8.50	82.70
x-vectors + DIHARD/RTVE PLDA + AHC	-	80.19	63.29	5.00	11.90	75.98
x-vectors + DIHARD/RTVE PLDA + AHC (oracle # speakers)	-	86.51	69.61	5.00	11.90	0.00
oracle SAD + PLDA mixture + AHC+ VBx	0.55, 0.40, 11, 0.80	15.86	14.56	1.30	0.00	34.89
x-vectors + PLDA mixture + AHC + VBx	0.55, 0.40, 11, 0.80	34.61	17.71	5.00	11.90	37.74
x-vectors + DIHARD/RTVE PLDA + AHC + VBx	0.10, 0.40, 11, 0.80	43.55	26.60	5.00	11.90	51.78
x-vectors + VoxCeleb PLDA + AHC + VBx	0.10, 0.40, 11, 0.80	**22.77**	**5.80**	5.00	11.90	**14.17**
x-vectors + PLDA mixture + AHC	-	32.74	15.80	5.00	11.90	76.20
x-vectors + PLDA mixture + AHC + VBx	0.10, 0.40, 11, 0.80	30.33	13.43	5.00	11.90	28.44

Table 3: *DER (%), speaker error (%), missed speaker (%), false alarm (%) and speaker number error (%) comparison of different setups for the RTVE test dataset (**post-submission** results are in bold letters).*

System	alpha, fa, fb, p	DER	SE	MS	FA	Speaker # error
x-vectors + DIHARD/RTVE PLDA + AHC	-	68.06	57.56	4.40	6.10	80.90
x-vectors + PLDA mixture + AHC + VBx	0.55, 0.40, 11, 0.80	39.48	29.08	4.30	6.10	49.82
x-vectors + DIHARD/RTVE PLDA + AHC + VBx	0.10, 0.40, 11, 0.80	36.03	25.60	4.30	6.10	48.50
x-vectors + VoxCeleb PLDA + AHC + VBx	0.10, 0.40, 11, 0.80	**27.63**	**17.20**	4.30	6.10	**27.07**
x-vectors + PLDA mixture + AHC	-	34.76	24.30	4.30	6.10	58.81
x-vectors + PLDA mixture + AHC + VBx	0.10, 0.40, 11, 0.80	32.69	22.29	4.30	6.10	37.90

ing; this was useful alongside DER during VBx parameter optimization.

The DER and speaker number error results of different setups for RTVE 2020 dev and test datasets are shown in Table 2 and Table 3, respectively. The initial part of our experiments followed a similar strategy to the Kaldi Callhome diarization recipe [20], for i-vectors, we used the RTVE 2018 dataset to train the extractor and PLDA models, and for x-vectors, we added VoxCeleb1+2, as the extractor model requires more data during training; then we tuned the AHC threshold with RTVE 2018 in order to use it to compute performance on the RTVE 2020 dev dataset. As shown in Table 2, the first x-vector-based system outperformed the i-vector-based one, which is expected, so we discarded further experimentation with i-vectors. It should be noted that, in both cases, the estimated number of speakers performed poorly, so we tested the AHC with an oracle number of speakers. Such a test was performed only for reference reasons, as in the final submission, the number of speakers per recording in the test dataset is unknown. We expected that the AHC with an oracle number of speakers would deliver better results, but it was not the case. We believe that unlike traditional speaker diarization datasets, the utterances in the RTVE datasets contained numerous speakers. Since the recordings are from TV broadcast shows, there is an imbalance of speaker corpus (e.j. RTVE 2020 test mean and standard deviation of speaker time: 345s and 838s, respectively). We tested using an energy SAD early in development, but its lousy performance in such conditions directly affected the DER performance; we immediately moved to the pre-trained TDNN SAD model, and it provided better performance by a large margin. Additionally, we provide results with an oracle SAD; despite the oracle segmentation, the speaker number inference error is almost the same, which indicates that it is due to the clustering strategy, as it underestimates how many speakers there are per utterance. We cannot blame the VBx aggressiveness of redundant speakers removal, as the pre- and post-submission experiments without it suffer the same underestimation problem. We believe the AHC may not be the best method in conditions with many speakers; further studies are required.

Table 3 presents the results of our submitted systems; the first one was our contrastive setup; its AHC threshold was calibrated using the RTVE 2020 dev dataset; we can see that it had similar results to its counterpart in Table 2. The second one was our primary system; its VBx parameters were manually calibrated using the RTVE 2020 dev dataset. Post-submission experiments show that our primary system could have performed better with a considerable change in the VBx alpha, obtaining a 6.79% absolute improvement in DER; furthermore, with the PLDA mixture's replacement with the VoxCeleb1+2 one, it obtains an additional 5.06% improvement. It is clear that our DIHARD II dev + RTVE 2018 PLDA hindered our mixture results; the addition of DIHARD II dev and heavy augmentations (reverberation and noises) most probably caused this.

4.1. Development resources

We conducted our experiments on the CLSP Cluster[3] on several Intel(R) Xeon(R) CPU E5-2680 v2 @ 2.80GHz nodes using up to 54 threads with 60 GB of RAM, and an NVIDIA GeForce GTX 1080 Ti with 11 GB of VRAM; Our submitted system took 35 hours for training and 15 hours to infer the RTVE test results.

5. Discussion

The x-vector-based system obtained our best results with an out-of-domain PLDA and VBx clustering. However, it falls behind

[3]http://www.statmt.org/jhu/?n=Info.CLSPCluster

our expected performance; we blame our heavy usage of out-of-domain-trained modules. Specifically, the x-vector extractor model, as previously mentioned, the RTVE dataset has specific peculiarities that differentiate it from standard speaker diarization datasets. The same can be said about the used TDNN SAD, as it was an out-of-the-box pre-trained model.

Our system would also be benefited from a better VBx parameter calibration, as shown in the *post-submission* results in Table 2 and Table 3. It should be noted that the usage of variational Bayes clustering greatly improved the system's ability to infer the number of speakers per recording, improving the DER.

6. Future work

Although we used a state-of-the-art approach for speaker diarization, we have room for improvements:

- We have to test domain-specific and hybrid approaches for the TDNN SAD model training, as its quality is directly associated with the diarization performance.
- In-domain data should be used for the x-vector extractor model training.
- Our system cannot handle overlapping speech, as it produces a single label per segment.
- Improve the system's speaker number inference in conditions such as the challenge's, where there are many speakers with imbalanced occurrences.

7. Conclusions

In this paper, we described our submission for the IberSPEECH-RTVE 2020 Speaker Diarization and Identity Assignment Challenge; we tested a state-of-the-art approach for diarization in a challenging Broadcast News scenario. We assessed the effectiveness of variational Bayes clustering as it significantly improved our system's ability to infer the number of speakers.

8. References

[1] D. Karim, C. Adnen, and H. Salah, "A system for speaker detection and tracking in audio broadcast news," in *2017 International Conference on Engineering MIS (ICEMIS)*, 2017, pp. 1–5.

[2] A. Ortega, A. Miguel, E. Lleida, V. Bazán, C. Pérez, M. Gómez, , and A. de Prada, "Albayzin Evaluation IberSPEECH-RTVE 2020 Speaker Diarization and Identity Assignment," 2020. [Online]. Available: http://catedrartve.unizar.es/reto2020/EvalPlan-SD-2020-v1.pdf

[3] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. de Prada, "RTVE2020 Database Description," 2020. [Online]. Available: http://catedrartve.unizar.es/reto2020/RTVE2020DB.pdf

[4] J. Patino, H. Delgado, R. Yin, H. Bredin, C. Barras, and N. Evans, "ODESSA at Albayzin Speaker Diarization Challenge 2018," in *Proc. IberSPEECH 2018*, 2018, pp. 211–215. [Online]. Available: http://dx.doi.org/10.21437/IberSPEECH.2018-43

[5] E. Lleida, A. Ortega, A. Miguel, V. Bazán, C. Pérez, M. Gómez, and A. de Prada, "RTVE2018 Database Description," 2018. [Online]. Available: http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf

[6] Z. Huang, L. P. García-Perera, J. Villalba, D. Povey, and N. Dehak, "JHU Diarization System Description," in *Proc. IberSPEECH 2018*, 2018, pp. 236–239. [Online]. Available: http://dx.doi.org/10.21437/IberSPEECH.2018-49

[7] M. Diez, F. Landini, L. Burget, J. Rohdin, A. Silnova, K. Žmolíková, O. Novotný, K. Veselý, O. Glembek, O. Plchot, L. Mošner, and P. Matějka, "BUT System for DIHARD Speech Diarization Challenge 2018," in *Proc. Interspeech 2018*, 2018, pp. 2798–2802. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1749

[8] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully Supervised Speaker Diarization," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6301–6305.

[9] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur, "Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge," in *Proc. Interspeech 2018*, 2018, pp. 2808–2812. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1893

[10] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," *Interspeech 2017*, Aug 2017. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-950

[11] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Interspeech 2018*, Sep 2018. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1929

[12] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The Second DIHARD Diarization Challenge: Dataset, task, and baselines," 2019.

[13] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015. [Online]. Available: http://arxiv.org/abs/1510.08484

[14] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "CHiME-6 Challenge:Tackling Multispeaker Speech Recognition for Unsegmented Recordings," 2020.

[15] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic Modelling from the Signal Domain Using CNNs," in *Interspeech 2016*, 2016, pp. 3434–3438. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-1495

[16] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 165–170.

[17] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[18] F. Landini, S. Wang, M. Diez, L. Burget, P. Matejka, K. Žmolíková, L. Mošner, O. Plchot, O. Novotny, H. Zeinali, and J. Rohdin, "BUT System Description for DIHARD Speech Diarization Challenge 2019," 10 2019.

[19] M. Diez, L. Burget, F. Landini, S. Wang, and H. Černocký, "Optimizing Bayesian Hmm Based X-Vector Clustering for the Second Dihard Speech Diarization Challenge," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6519–6523.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Vesel, "The Kaldi speech recognition toolkit," *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 01 2011.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Diarization and Identity Assignment Compatibility in the Albayzín 2020 Challenge

Ignacio Viñals, Pablo Gimeno, Alfonso Ortega, Antonio Miguel and Eduardo Lleida

University of Zaragoza, Spain
{ivinalsb, pablogj, ortega, amiguel, lleida}@unizar.es

Abstract

The current need to identify the speakers in a certain recording has evolved along time, requesting more and more information. While speaker recognition originally focused on determining whether a speaker talks in a certain audio with a single speaker, later diarization focused on differentiating speakers along the recording. The latest step is Identity Assignment (IA), which combines both of them, i.e., deciding whether a certain speaker is present in a given audio, as well as determining the periods of time when the speaker is active.

Our work presents and analyzes the ViVoLAB results for the Albayzín 2020 evaluation, focused on diarization and identity assignment. These challenges will be faced in the broadcast domain, with data coming from national Spanish TV Corporation RTVE. For this purpose we have developed a Bottom-Up diarization architecture based on the embedding-PLDA paradigm. On top of the diarization solution we have added an identity assignment block, based on the speaker verification approach.

Index Terms: diarization, identity assignment, neural networks.

1. Introduction

The recent increase of raw multimedia data has made capital the development of automatic techniques capable of labeling any input audio. Among these techniques diarization is the task dedicated to identify the time stamps defining when any speaker talks in a given audio. Diarization goal is the differentiation of the speakers rather than their real identification. For this purpose we can make use of generic labels.

Diarization has evolved simultaneously to other speech technologies along time, specially inheriting approaches coming from speaker recognition: From basic technologies [1] to Joint Factor Analysis (JFA) [2, 3], i-vectors [4] or Deep Neural Networks (DNNs) [5]. Furthermore, this development also allows us to obtain reasonable performances out of telephone domain, as in broadcast [6] or meetings domains [7].

Nevertheless, these labels become worthier as long as they become more and more informative. Thus, the generic labels obtained during diarization could be improved if the true identity of the speaker was inferred too. This task, also known as Identity Assignment (IA), can be considered a complementary task working on top of diarization [3]. Although diarization was a complementary task for IA during its early days, now it has gained more and more relevance as long as we need higher and higher quality information from an audio. The joint work of both blocks helps us determining whether a certain speaker is present in a given audio as well as determining when he is contributing.

Albayzín 2020 is the most recent edition in the ongoing series of Albayzín technological evaluations, seeking the improvement of speech technologies in Iberian languages (languages spoken in the Iberian peninsula), paying special attention on the broadcast domain. 2020 edition continues the line from previous evaluations, working with audios provided by Radio Televisión Española (RTVE), the national Spanish TV corporation. However, in addition to diarization, 2020 edition now includes the Identity Assignment problem as a complementary goal.

In this work we present the ViVoLAB submission to Albayzín 2020 evaluation, specifically to the diarization and identity assignment task. Our system considers the dual problem as a cascade of tasks. First we perform diarization considering a Bottom-Up approach where the original audio, once processed by the front-end, is segmented and then clustered to obtain the final speaker labels. Regarding the IA task, we manage it as a speaker verification problem where diarization clusters play the role of test audios to evaluate against the given enrollment recordings.

The rest of the document is structured as follows: In Section 2 we describe the ViVoLAB diarization system. The identity assignment block is explained in Section 3. Our experimental results are included in Section 4. Finally, our conclusions are expressed in Section 5.

2. ViVoLAB diarization system

The diarization system works according to the Bottom-Up philosophy, i.e., first identifying segments with a single speaker on them, later combined according to a clustering block. This clustering block makes use of the embedding-PLDA (Probabilistic Linear Discriminant Analysis) paradigm.

2.1. Voice Activity Detection

Our approach for voice activity detection (VAD) is based on a deep learning solution. We use a convolutional recurrent neural network (CRNN) consisting of 3 2D convolutional blocks (2D conv. layer with 64 filters of size 3x3, batch normalization and ReLU activation) followed by 3 Bidirectional Long Short Time Memory (BiLSTM) layers. Then, the final speech score is obtained through a linear layer. The neural network works in terms of streams of feature vectors, 300 to be specific, inferring a VAD label per feature in the input sequence. As input features, 64 Mel filter banks and the frame energy are extracted from the raw audio and fed to the neural network. These input features are normalized in mean and variance prior to any other calculation within the network.

Adam is chosen as the optimizer for the neural network, using a learning rate that decays exponentially from 10^{-3} to 10^{-4} in the 30 epochs that data is presented to the neural network. Cross entropy is the training objective, as usually done in classification tasks.

10.21437/IberSPEECH.2021-20

The CRNN is trained on a combination of different broadcast datasets. Specifically, we include data from the Albayzín 2010 dataset [8] (train and eval), Albayzín 2018 dataset [9] (dev2 and eval) and a selection of data from 2015 Multi-Genre Broadcast (MGB) Challenge [10] (train, dev. longitudinal and task3 eval). A 10% of all the data is reserved for training validation. Furthermore, audios are augmented with a variety of noises that can be usually found in broadcast emissions (sitcom noises, crowd and laughter noises, babble, street music and stadium noises). These noises are added in training time with a Signal to Noise ratio (SNR) that is sampled from an uniform distribution in the range (5, 25) dB.

2.2. Speaker Change Point Detection

The Speaker Change Point Detection block works in terms of Bayesian Information Criterion (BIC), according to its differential form (ΔBIC) [11]. We consider analysis windows of 6 seconds, modelling speakers with full- covariance Gaussian distributions. This block prioritizes those speech/non-speech boundaries given by VAD. As input features the system considers 20 MFCC [12] features vectors, over a 25 ms hamming window every 10 ms. Features are then normalized according to Cepstral Mean and Variance Normalization [13] to mitigate channel effects.

2.3. Embedding Extraction

Each one of the obtained segments will be transformed into a compact representation also known as embedding. For this purpose we have opted for an evolution of the extended x-vector [14] architecture, based on Time Delay Neural Networks (TDNNs). Compared to the original architecture, we have substituted the mean and standard deviation pooling block by a multi-head self-attention block [15]. This self-attention block simultaneously considers H different patterns to learn from the data themselves, also known as heads. For each pattern j this block weighs the stream of forwarded information $\mathbf{X} = \{\mathbf{x}_1, .., \mathbf{x}_i, ..\mathbf{x}_N\}$ by a set of weights $\boldsymbol{\alpha}_{ij}$ estimated as follows:

$$\boldsymbol{\alpha}_{ij} = \frac{\exp\left(\mathbf{W}_j \mathbf{x}_i + \boldsymbol{b_j}\right)}{\sum_i \exp\left(\mathbf{W}_j \mathbf{x}_i + \boldsymbol{b_j}\right)} \tag{1}$$

where $\mathbf{W}_j$ refers to the row j of the weight matrix $\mathbf{W}$, and $\boldsymbol{b}_j$ stands for the jth component of the bias vector $\boldsymbol{b}$. Both weight matrix and bias are the trainable parameters of an affine transformation of H neurons.

Given the obtained weights $\boldsymbol{\alpha}_{ij}$, the block estimates for each head j its weighted mean ($\boldsymbol{\mu}_j$) and weighted standard deviation ($\boldsymbol{\sigma}_j$). The output of the block consists of the concatenation of the estimated mean and standard deviation from each head. Experimentally we set the value of H up to 8. The final configuration for the network is shown in Table 1.

The neural network has been trained with data from VoxCeleb 1 [16] and 2 [17] The resulting neural network provides embeddings of dimension 512. These embeddings will be later centered, dimensionsionally reduced by means of LDA up to 200 as whitening and length-normalized [18].

2.4. Clustering

The obtained embeddings are modeled in a generative manner according to [19], where a tree-based PLDA clustering is proposed. This solution proposes a Maximum A Posteriori (MAP) estimation of the speaker labels Θ given the set of embeddings

#	Component Type	Context	Size
1	TDNN-ReLU-BN	[t-2:t+2]	512
2	FFN-ReLU-BN	t	512
3	TDNN-ReLU-BN	[t-2,t,t+2]	512
4	FFN-ReLU-BN	t	512
5	TDNN-ReLU-BN	[t-3,t,t+3]	512
6	FFN-ReLU-BN	t	512
7	TDNN-ReLU-BN	[t-4,t,t+4]	512
8	FFN-ReLU-BN	t	512
9	FFN-ReLU-BN	t	1536
10	Multi-Head Self-Att. Pool.	Full Seq.	H*1536*2
11	FFN-ReLU-BN	Full Seq.	512
12	FFN-ReLU-BN	Full Seq.	512
13	Softmax	Full Seq.	N_{spk}

Table 1: *Architecture for the embedding extractor. Involved elements are Time Delay Neural Network (TDNN), Rectified Linear Unit (ReLU), Batch Normalization (BN) and Feed Forward Network (FFN). Context explains which frames at the input of the layer are taken into account to build the t-th output frame. Layers with full sequence (Full Seq.) context work at utterance level.*

$\boldsymbol{\Phi}$ in order to obtain the diarization labels Θ_{DIAR}:

$$\Theta_{\mathrm{DIAR}} = \arg\max_{\theta} P\left(\Theta|\boldsymbol{\Phi}\right) = \arg\max_{\theta} P\left(\boldsymbol{\Phi}|\Theta\right) P\left(\Theta\right) \tag{2}$$

The model considers a Fully Bayesian PLDA [20] of dimension 100 to explain $P\left(\boldsymbol{\Phi}|\Theta\right)$, while the priors $P\left(\Theta\right)$ follow [21] making use of a modification of the Distance Dependent Chinese Restaurant (ddCR) process. Additionally, we interpret $P\left(\boldsymbol{\Phi}, \Theta\right)$ as a tree structure by means of the product rule of probability. Hence, we opt for an optimization of the model according to a sequential manner making use of the M-algorithm [22] to find the best possible path along the tree. Moreover, prior to any clustering evaluation, the PLDA model is adapted thanks to unsupervised adaptation approaches as described in [6].

3. Identity Assignment

The Identity Assignment (IA) block in ViVoLAB submission follows the schematic of a speaker verification task based on the standard embedding-PLDA paradigm. Hence, as preparation each one of the enrollment recordings is converted into its corresponding embedding as well as the obtained segments from diarization. For the speaker verification evaluation, enrollment models are built according to the corresponding given audios while test models represent the clusters obtained during diarization. Each test model is made in terms of all segments assigned to the cluster. For simplicity reasons we make use of the same embedding extractor and PLDA trained for diarization purposes.

The obtained scores are then normalized by means of adaptive S-normalization

$$s' = \frac{s - \mu_t}{\sigma_t} + \frac{s - \mu_e}{\sigma_e} \tag{3}$$

where the score s is transformed into the score s' in terms of the the means μ_t and μ_e and standard deviations σ_t and σ_e. While μ_t and σ_t are computed on the scores of the cohort versus the test segments, μ_e and σ_e are computed on the scores of

Table 2: *Results of the ViVoLAB system for the two subtasks, diarization and identity assignment. Results obtained for two subsets, development and test.*

Metric	Results (%)	
	Dev.	Test
Diarization		
DER	16.96	15.24
Identity Assignment		
AER	47.90	72.63
ASE	128.11	505.82

the enrollment segments versus the cohort. The chosen normalization cohort in our experiments consisted of the MGB 2015 dataset. The score normalization was adaptive. For each segment, we select cohorts similar to the test segment to compute the normalization values. For each trial, we selected 25% of the total segments in the cohort. The selection is based on the own PLDA scores.

The final labels are built according to a threshold adjusted during calibration. This adjustment was obtained experimentally with the development set. Furthermore, as a design choice we do not exclude the possibility of multiple clusters assigned to the same enrollment. This decision was made in order to allow the correction of diarization errors.

4. Results

Due to the fact that 2020 evaluation considers two different tasks, we need metrics for both of them. Regarding diarization, the evaluation takes into account the commonly used Diarization Error Rate (DER) , which determines the ratio of mislabeled audio to the total audio to analyze. With respect to identity assignment two metrics were proposed: On the one hand we have Assignment Error Rate (AER), similar to DER, although only matching clusters from hypothesis and reference when sharing the same speaker label. On the other hand we have Average Speaker Error (ASE), an average of the ratio of error per speaker along the subset of interest. Once introduced the three metrics, the results for ViVoLAB system are shown in Table 2.

The results given in Table 2 evidence multiple trends. First, diarization results offer a good performance, not suffering from great mismatches between development and test. This is very important when involved shows within both subsets may not match. Besides, these results follow the trend from previous evaluations [9], specially considering the addition of more complex audio. With respect to Identity Assignment results, we observe a high degradation compared to DER results in both types of metric (AER and ASE). In fact, this degradation is much more severe in the second metric (ASE). Moreover, this degradation does not affect development and test in a similar way but specially harms evaluation scores.

The first analysis of interest studies the decomposition of DER into its three decoupled terms: miss (speech not considered to contain human voice), false alarm (audio mistakenly labeled to contain speech) and speaker (speech misclassified among the speakers). While the first two are related to the VAD stage, the latest one is only influenced by the SCPD as well as clustering. The results for this analysis are illustrated in Table 3:

Table 3: *Decomposition of DER into its three terms, Miss, False Alarm (F.A.) and Speaker (SPK). Analysis performed for both development and test subsets.*

Subset	Errors(%)		
	MISS	F.A.	SPK
Development	3.61	2.74	10.61
Test	3.65	1.97	9.62

Table 4: *Decomposition of AER in False Alarm (F.A.), Miss (MISS) and Speaker (SPK) errors for development and test subsets*

Subset	Errors(%)		
	MISS	F.A.	SPK
DEV	14.0	29.6	4.3
TEST	5.2	57.0	14.5

The results in Table 3 evidence a reasonable good performance of the VAD block, being the posterior blocks responsible for most of the diarization error. This VAD performance is specially relevant when bearing in mind that its labels work as anchors for posterior stages and its errors cannot be compensated afterwards. Moreover we also want to highlight the robustness of this VAD block, working similarly with both subsets thanks to its generalization capabilities.

Apart from the traditional diarization subtask, the interest for identity assignment as well as its poor results encourages to take a deeper look into the new subtask. Our first analysis is a decomposition of AER into its three composite terms:

- E_{MISS}, or miss error, determines how much speech is lost.
- E_{FA}, also known as false alarm error, illustrates how much non-desired audio is considered as speech.
- E_{SPK}, named as speaker error, indicates how much speech is mistakenly assigned among the speakers.

This analysis is carried out for both development and test subsets. The obtained results are included in Table 4.

According to those results shown given in Table 4 we can conclude that the main cause of error is the False Alarm term. In fact, this error implies at least a relative 60% of the whole error in both subsets. Consequently too much undesired audio is considered as coming from the target speakers. This error also explains the high values for ASE, highly affecting speakers of interest with limited speech contributions. Besides, the differences in this error term between development and test subsets are responsible for the great difference in performance between both subsets.

Nevertheless, we still must take into account the way ViVoLAB system faced the IA task, i.e. by means of a speaker verification evaluation. In this evaluation enrollment models are created according to the given audios while test models are built according to diarization labels. Our next analysis studies DET curves for both development and test subsets. These curves are illustrated in Fig. 1, also including the EER (Equal Error Rate) value. Please notice that this analysis omits the amount of speech contained in each cluster, treating them evenly for scoring purposes.

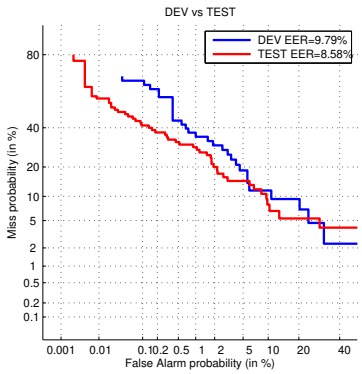


Figure 1: *DET curves for development (blue) and test (red) subsets, also indicating the EER.*

According to the curves shown in Fig. 1 we see high degradation values for both subsets. Moreover, along the whole curves the test subset is always outperforming its development counterpart regardless of the operation point. These results seem to disagree with the previously obtained results.

Our final study consists of analyzing the score distribution per subset, development and test, as well as type of trial, target and non-target. This analysis is carried out considering score histograms. In Fig. 2 we reproduce our analysis, illustrating in thick and dashed lines the scores for non-target and target trials respectively. In the same figure we have included the scores for both development (blue line) and test (red line) subsets.

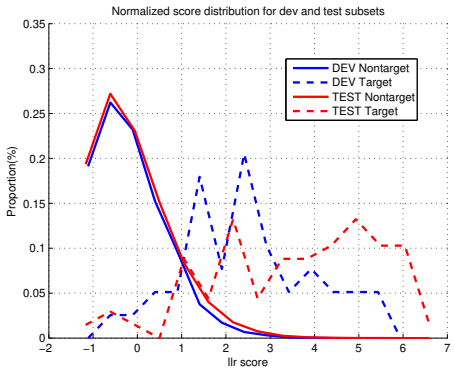


Figure 2: *Score histograms for development (blue) and test (red) subsets, differentiating between non-target (thick lines) and target (dashed lines) trials*

The histograms in Fig. 2 explain the behaviour of the system. While non-target trials offer similar score distributions for development and test subsets, target trials do not. Target distributions for development and test subsets are slightly shifted from each other. Hence, the calibration task performed during development is not valid anymore with test data. This circumstance also explains why a better DET curve provides worse IA results. This mismatch may have many reasons, such as quality of the clusters as well as the domain mismatch between enrollment and test data. This latest factor is specially relevant when comparing diarization with IA tasks. While diarization does only deal with a single domain, given by the test audio, IA may have few of them, including the test audio as well as those present in the enrollment data. In addition to this factor IA must also fit a threshold, ideally robust against any of these mismatches.

5. Conclusions

Along the previous lines we have seen the performance of ViVoLAB system in two different tasks, diarization and identity assignment. While its performance in diarization seems promising with good results, we notice a great degradation when evaluating IA.

Diarization results follow the trend of performance from previous evaluations despite the higher complexity of 2020 data. Moreover, unsupervised adaptation techniques help minimizing the mismatch between development and test subsets, offering similar results.

By contrast, IA results suffer from a significant degradation. According to the carried out analysis we see that the main source of degradation is the large amount of false alarm errors, i.e. undesired speakers considered as part of the target ones. This type of degradation specially influences ASE metric. In addition to this factor, the speaker verification treatment of the I problem suffers from mismatch in score histograms between enrollment and test, causing a calibration issue. This can be partially caused by the quality of clusters, given by diarization, as well as acoustic domain mismatches of data, both enrollment and test. Regardless of the cause, its solution is capital for the robustness of this type of systems and its addition to real world applications.

6. References

[1] S. E. Tranter and D. Reynolds, "An Overview of Automatic Speaker Diarization Systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[2] F. Valente, P. Motlicek, and D. Vijayasenan, "Variational Bayesian Speaker Diarization of Meeting Recordings," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4954–4957.

[3] C. Vaquero, A. Ortega, A. Miguel, and E. Lleida, "Quality Assessment of Speaker Diarization for Speaker Characterization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 4, pp. 816–827, 2013.

[4] J. Villalba, A. Ortega, A. Miguel, and E. Lleida, "Variational Bayesian PLDA for Speaker Diarization in the MGB Challenge," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 667–674.

[5] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker Diarization Using Deep Neural Network Embeddings," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4930 – 4934.

[6] I. Viñals, A. Ortega, J. Villalba, A. Miguel, and E. Lleida, "Unsupervised adaptation of PLDA models for broadcast diarization," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, 2019.

[7] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, L. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI Meeting Corpus: A pre-announcement," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3869 LNCS, pp. 28–39, 2006.

[8] T. Butko and C. Nadeu, "Audio segmentation of broadcast news in the albayzin-2010 evaluation: overview, results, and discussion,"

EURASIP Journal on Audio, Speech, and Music Processing, vol. 2011, no. 1, pp. 1–10, 2011.

[9] E. Lleida, A. Ortega, A. Miguel, V. Bazán, C. Pérez, M. Gómez, and A. de Prada, "Albayzin 2018 evaluation: The IberSpeech-RTVE challenge on speech technologies for Spanish broadcast media," *Applied Sciences*, vol. 9, no. 24, pp. 1–22, 2019.

[10] P. Bell, M. J. F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, and P. C. Woodland, "The MGB Challenge: Evaluating Multi-Genre Broadcast Media Recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 687–693.

[11] S. S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering Via the Bayesian Information Criterion," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, vol. 6, pp. 127–132, 1998.

[12] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[13] M. J. Alam, P. Ouellet, P. Kenny, and D. O'Shaughnessy, "Comparative evaluation of feature normalization techniques for speaker verification," *Advances in Nonlinear Speech Processing. NOLISP 2011. Lecture Notes in Computer Science*, vol. 7015, no. 2011, pp. 246–253, 2011.

[14] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, F. Richardson, S. Shon, F. Grondin, R. Dehak, L. P. Garcia-Perera, D. Povey, P. Torres-Carrasquillo, S. Khudanpur, and N. Dehak, "State-of-the-art Speaker Recognition for Telephone and Video Speech: the JHU-MIT Submission for NIST SRE18," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1488–1492, 2019.

[15] D. Bahdanau, K. Cho, and Y. Bengio, "NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE," in *International Conference on Learning Representations (ICLR)*, 2015, pp. 1–15.

[16] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-Augus, pp. 2616–2620, 2017.

[17] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep Speaker Recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-Septe, no. ii, pp. 1086–1090, 2018.

[18] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2011, pp. 249–252.

[19] I. Viñals, P. Gimeno, A. Ortega, A. Miguel, and E. Lleida, "ViVoLAB Speaker Diarization System for the DIHARD 2019 Challenge," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 988–992, 2019.

[20] J. Villalba and E. Lleida, "Unsupervised Adaptation of PLDA By Using Variational Bayes Methods," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 744–748.

[21] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully Supervised Speaker Diarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6301–6305.

[22] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

The Biometric Vox System for the Albayzin-RTVE 2020 Speech-to-Text Challenge

Roberto Font[1], *Teresa Grau*[1]

[1]Biometric Vox S.L.

roberto.font@biometricvox.com, teresa.grau@biometricvox.com

Abstract

This paper describes the system developed by Biometric Vox for the Albayzin Speech-To-Text Challenge organized as part of the Iberspeech 2020 conference. The system uses speaker diarization to segment the audio into speaker-homogeneous segments and uses this information to compute speaker-dependent fMLLR transformed features. These speaker-adapted features are the input to a DNN acoustic model which is trained for the domain at hand using a semi-supervised self-training procedure. Finally, a RNN language model is used for lattice rescoring and producing the final transcription. Our system achieves 22% WER on the test portion of the RTVE2018 database and 30,26% on the 2020 evaluation set.

Index Terms: speech recognition, Hybrid DNN-HMM, semi-supervised training, self-training

1. Introduction

The aim of this paper is to provide a complete description of the system developed by Biometric Vox for the Albayzin-RTVE 2020 Speech-to-Text Challenge. The task for this challenge is the automatic transcription of TV shows that belong to diverse genres and broadcast by the public Spanish Television (RTVE), covering a wide range of realistic and challenging conditions: spontaneous and read speech, different accents, background noise, songs, laughs, yells, quick short-turn conversations...

Our main contribution is the use of a very simple yet effective semi-supervised learning method for acoustic model training. Starting from a commercial off-the-shelf system that was developed for the automatic transcription of town hall plenaries, we use self-training to adapt this system to the domain at hand without any human intervention. The transcriptions produced by this initial system are used to train a new system and this procedure can be repeated iteratively. For this work, we performed two of these self-training iterations.

The rest of the paper is organized as follows: in Section 2 the RTVE database and the data used to train the acoustic model and language models are described. Section 3 presents the ASR system and describes its components and characteristics. Results are shown in Section 4, and, finally, we summarize our results and conclusions in Section 5.

2. Data Resources

2.1. RTVE database

The RTVE2018 database [1] has a total of 569 hours and 22 minutes of audio extracted from 17 different TV shows broadcast by RTVE (Radio Televisión Española) from 2015 to 2018. Most shows are related to news, debates, social gatherings and documentaries. The database is divided into 4 subsets: a training set, 2 development sets, *dev1* and *dev2*, and a test set, as summarized in Table 1.

Table 1: *Composition of the different RTVE subsets*

Subset	# hours	# different shows
RTVE2018		
train	500h	16
dev1	52:31:51	5
dev2	15:09:25	2
test	39:07:15	8
RTVE2020		
dev	03:55:31	2
test	67:23:29	15

It is worth noting that the *dev1*, *dev2* and *test* portions have human-revised transcriptions while, in the case of the *train* set, it contains subtitles that were generated through a re-speaking procedure resulting in a no-verbatim word transcription.

Additionally, the database includes a text corpora extracted from all the subtitles broadcast by the RTVE 24H Channel during 2017. The subtitles contain approximately 56M words.

The RTVE2020 database [2] is also a collection of TV shows from the public Spanish Television (RTVE) that were broadcast from 2018 to 2019. The database is composed of 70 hours and 18 minutes of audio belonging to 15 different shows. Only two shows are in common with the 2018 dataset: *Comando Actualidad* and *Millenium*. For the challenge, no transcriptions or subtitles were provided for this set. The database is divided into 2 subsets: *dev* and *test*. The latter one is the challenge evaluation set. Shows' genre for this set is related to entertainment: series, sketches, reports, entertainment shows...

TV show's genre and format are quite different between the two sets; each set covers a different range of conditions. For the RTVE2018 database: spontaneous and read speech, long-turn conversations, different accents and background noise. And for RTVE2020: songs, laughs, yells, quick short-turn conversations and colloquial vocabulary.

For system development, we used the *dev1* portion of the RTVE2018 database as our internal development dataset and the RTVE2018 *test* portion as our test set.

2.2. Hybrid (DNN-HMM) acoustic model

As already discussed, the training portion of the RTVE2018 database contains imprecise transcriptions, since the subtitles have been generated by a re-speaking procedure resulting in a

[1]http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf

[2]http://catedrartve.unizar.es/reto2020/RTVE2020DB.pdf

10.21437/IberSPEECH.2021-21

no verbatim word transcription. This prevents the usage of this material in a standard acoustic model training setup.

To account for this difficulty, we have used a semi-supervised learning method that was developed for our commercial automatic transcription system *Transcribe Vox* [3]. This system, which automatically transcribes town hall plenaries and similar meetings, was trained using an iterative self-training procedure. First, an initial system was bootstrapped from a small amount of labelled data, then this initial system was used to produce transcriptions on a large amount of unlabelled data. These transcriptions were used to train a new system and this procedure was repeated iteratively to produce increasingly accurate transcriptions.

For the Albayzin-RTVE 2020 Speech-to-Text Challenge we followed the same approach using this pre-existing model to generate transcriptions for the training portion of the RTVE2018 database. These transcriptions were used, instead of the provided subtitles, to train our system.

To sum up, the acoustic model is trained on:

- A set of roughly 530 hours of town hall meetings and the transcriptions generated by the initial system. This material, publicly available, was downloaded from the Internet trough the different platforms offered to citizens to review the meetings.
- RTVE2018 training set and the transcriptions generated by the initial system (trained on the above set).

To increase model robustness, we applied data augmentation for DNN training. We produced volume and speed perturbations and added reverberation and noise using the Musan corpus [1]. A subset of the augmented data was selected obtaining a final training set of approximately $1,800$ hours.

2.3. Language modelling

To train the LMs the following text corpora were used:

In-domain data: The RTVE subtitles provided in the RTVE2018 database. The subtitles contain approximately 56M words.

Out-of-domain data: The combination of the Spanish portion of the Europarl corpus [2] and the automatically-generated transcriptions of the town hall meetings described in the previous section, with a total of around 60M words.

The text was preprocessed by the usual procedure of normalizing, removing punctuation, expanding the most usual contractions and transliterating numbers. We used a vocabulary of 101K words including the 320 most frequent words in the RTVE subtitles that were not present in the off-the-shelf lexicon.

Two kinds of LMs have been trained:

- **3-gram LM:** A trigram language model obtained by linear interpolation of two models: one trained on in-domain data, and the other on the out-of-domain dataset. The optimal interpolation weight was tuned on the development set.
- **RNNLM:** An RNN language model trained on the combination of both datasets.

[3] https://biometricvox.com/transcribevox/

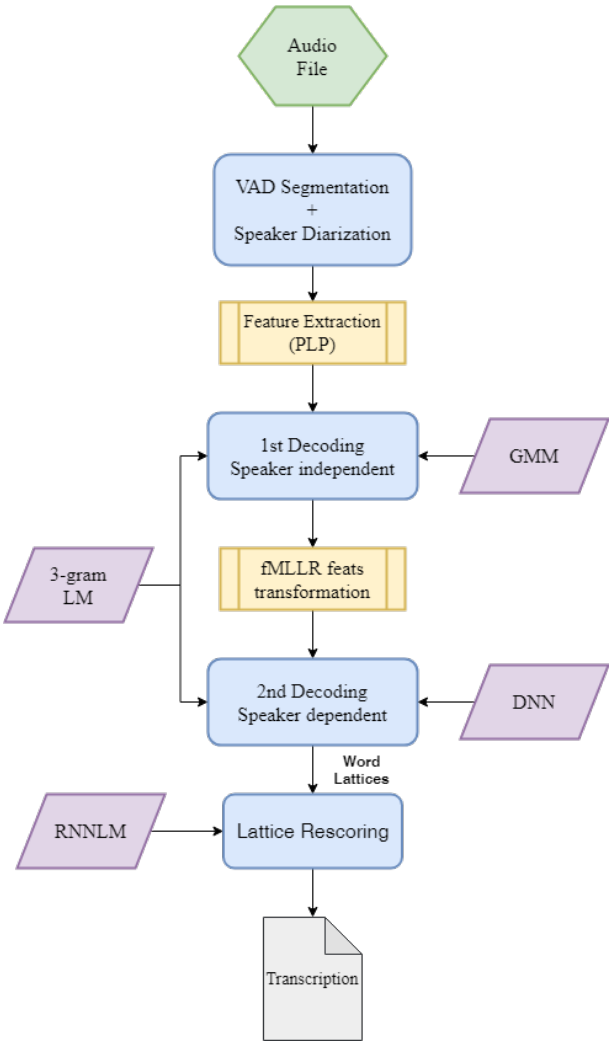


Figure 1: *System flowchart*

3. System Components

The transcription process is composed of the following steps, which were implemented using the Kaldi [3] toolkit:

- Initial segmentation and silence removal using Kaldi standard energy-based VAD.
- Speaker diarization to merge together the segments belonging to the same speaker.
- First-pass speaker independent decoding using a triphone GMM system. This first-pass decoding is used to compute fMLLR transforms.
- Second-pass decoding using fMLLR-transformed features and a DNN model.
- Lattice rescoring [4] using an RNN language model.

Figure 1 shows an schematic view of this process.

3.1. Speaker diarization

To merge together all segments belonging to the same speaker and compute speaker-adapted features, we perform a speaker

diarization step using the system developed for the Albayzin-RTVE 2020 Speaker Diarization and Identity Assignment Challenge. The system is based on the DNN x–vector paradigm [5] and consists of the following steps:

- Acoustic feature extraction (23 MFCC features) and Voice Activity Detection using an energy-based VAD.
- X–vector embedding extraction.
- Embedding post-processing (length-normalization, centering, whitening, LDA)
- PLDA scoring.
- Agglomerative Hierarchical Clustering.

The embedding extractor is trained on NIST SRE 04-10, MIXER6, Switchboard and VoxcelebCat. VoxcelebCat is the result of concatenating all excerpts from the same video into one longer file and combining Voxceleb 1 train, Voxceleb 2 dev and Voxceleb 2 test. The samples from NIST SRE 04-10, MIXER6 and Switchboard were upsampled to 16kHz. As is usual, we augment the training data by generating perturbed versions using Musan [1] noise and reverbation.

For the embedding extractor, we used a baseline TDNN x-vector architecture as in the Kaldi SRE 16 recipe (Table 2).

Table 2: *Embedding extraction architecture for speaker diarization*

Layer type	**Layer context**	**Size**
TDNN-ReLU-batchnorm	t-2:t+2	512
TDNN-ReLU-batchnorm	t-2, t, t+2	512
TDNN-ReLU-batchnorm	t-3, t, t+3	512
ReLU-batchnorm	t	512
ReLU-batchnorm	t	1500
Stats Pooling (mean+stddev)	T	2x1500
ReLU-batchnorm		512
ReLU-batchnorm		512
Softmax		# speakers

Embeddings are extracted using a sliding window of 3 seconds and a second and a half hop. Then, they are length-normalized, projected from 512 dimensions to 150 using LDA and scored using PLDA. Finally, segments are clustered using Agglomerative Hierarchical Clustering (AHC).

Both LDA and PLDA are trained on VoxcelebCat.

3.2. First-pass GMM acoustic model

The GMM model used in the first-pass speaker independent decoding is an LDA+MLLT+SAT triphone system with $4,360$ classes (tied-state triphones) and $60,094$ gaussians that has as input 13-dimensional PLP features spliced across 7 consecutive frames and projected to 40 dimensions trough the LDA+MLLT transformation. The training set for this model is the same described in Section 2.2 except for the augmented data, that was not included. As already mentioned, this first-pass decoding is used to compute the fMLLR-transformed features, the input for the DNN acoustic model.

3.3. DNN acoustic model

The DNN acoustic model architecture is based on Kaldi's *chain* model [6]. A factorized time-delay neural network (TDNN-F) [7], which is structurally the same as a TDNN [8], but is trained from a random start with one of the two factors of each matrix constrained to be semi-orthogonal.

Our network consists of 7 tdnn-f layers with dimension $1,536$ and linear bottleneck layers of dimension 256, ReLU activation function and batch normalization. For the loss function we used L2 regularization to prevent overfitting. Table 3 shows the architecture of the tdnnf. Instead of using ivectors for speaker adaptation, as is often the case, we use fMLLR-transformed features as input for the DNN. These transformations are computed using the first-pass decoder described above. To improve the accuracy of this speaker adaptation, we do an initial speaker diarization step to try and group together all segments belonging to the same speaker.

Table 3: *Factorized TDNN architecture. Note that Batch-Norm applied after each ReLU is omitted for brevity.*

Layer type	**Context factor1**	**Context factor2**	**Size**	**Inner size**
tdnn-ReLU	t		1536	
tdnnf-ReLU	t-1, t	t, t+1	1536	160
tdnnf-ReLU	t-1, t	t, t+1	1536	160
tdnnf-ReLU	t-1, t	t, t+1	1536	160
tdnnf-ReLU	t	t	1536	160
tdnnf-ReLU	t-3, t	t, t+3	1536	160
tdnnf-ReLU	t-3, t	t, t+3	1536	160
tdnnf-ReLU	t-3, t	t, t+3	1536	160
Linear			256	
Dense-ReLU-Linear			256	1536
Dense			#Targets	

TDNN training time for 4 epochs and 1763 iterations was 40 hours 52 minutes on an NVIDIA Geforce GTX 1080 GPU.

3.4. Language models

As mentioned in Section 2.3, we used a 3-gram LM which is the linear interpolation of in-domain and out-of-domain LMs for decoding and a Recurrent Neural Network LM for the final lattice rescoring.

For the 3-gram LM, the models were trained using the SRI Language Modeling Toolkit [9], and the optimal interpolation weight was tuned on the development set (*dev1*). Table 4 shows the perplexities for the out-of-domain, in-domain and interpolated LMs.

Table 4: *Language Model perplexity*

LM	**Perplexity**
Out-of-domain	440.65
In-domain	235.29
Interpolation	205.36

The RNNLM was also trained using the Kaldi toolkit. The network is trained with an architecture of 5 hidden layers, each with 800 neurons, where TDNN layers with ReLU activation function, and LSTM layers are combined. Figure 2 shows the topology of the network used.

The RNNLM was trained for 480 iterations, being the 478^{th} the best iteration. 10K sentences from the train set were used as development set.

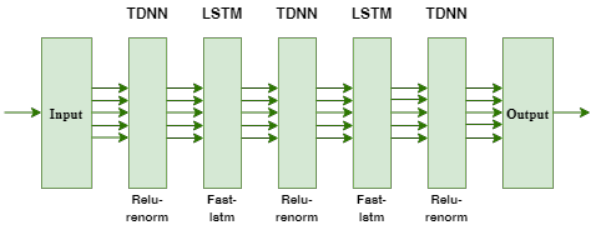


Figure 2: *RNNLM topology*

4. Results

The Table 5 shows the results of our system on the development and test portions of the RTVE2018 database.

Three systems are evaluated. The first one is the off-the-shelf commercial system described in Section 2.2. The other two systems are the result of the first and second self-training iterations as described in Section 2.2. No more self-training iterations were made due to time constraints. "RTVE2.0" system was our primary submission for the challenge. The evaluation results provided by the organizers of our primary system for the 2020 S2T Challenge's evaluation set are shown in Table 6. The table shows the total number of words and the average Word Error Rate (WER) by TV show, and also the global average WER.

We can see that the best result is obtained for the show *Los desayunos de TVE* (LD), which is related to news content and long-turn conversation almost without interruptions between speakers. On the other hand, the highest WER is obtained for the shows *Como nos Reíamos* (CN) and *Si Fueras Tú* (SFT), which are entertainment shows with sketches with a loud voice tone and loud background music.

The decoding was performed on an Intel(R) Core(TM) i7-5820K CPU @ 3,30GHz with 6 cores, and the processing time for the *dev2* subset was 4 hours and 27 minutes.

Table 5: *WER (%) on the RTVE2018 dataset*

System	dev2	Test
Off-the-shelf system	21.8%	-
RTVE1.0 + RNNLM_RTVE	20.3%	23.6%
RTVE2.0 + RNNLM_RTVE	17.8%	22.0%

5. Conclusions

We have presented our ASR system submitted to the Albayzin Speech-To-Text Challenge. This challenge, which focuses on the automatic transcription of TV shows, provides researchers with an ASR task with some very interesting and challenging features. Indeed, the provided training data contains imprecise transcriptions making it difficult to use it in a standard acoustic model training setup. In addition, the TV shows include some of the most challenging conditions for any speech recognition system: spontaneous speech, different accents, noisy backgrounds, overlapped speakers... Furthermore, those conditions are different between the RTVE2018 and RTVE2020 database. The first one focuses mainly in news content, where the speech tends to be slow and with long-turn conversations. Meanwhile, RTVE2020 focuses on entertainment shows with songs, sketches with a loud voice tone, quick short-turn conversations...

Table 6: *Evaluation results RTVE2020. #W column is for the total number of words and %WER for the average Word Error Rate by TV show.*

TV Show	#W	(%)WER
AT	108771	26.25
BN	6224	55.16
BR	6139	50.68
CA	37942	36.27
CN	20628	69.20
EP	21402	34.58
IM	24474	49.04
LD	121067	13.86
MC	85464	35.23
ML	16949	23.33
NFMY	1433	45.57
SFT	2108	54.87
VC	37458	36.32
VE	26314	24.29
WU	3406	52.17
Global	519779	30.26

Our main contribution is the use of a semi-supervised self-learning method to train the system without the need of labelled data for the domain at hand. The initial system, trained on data from town hall plenary sessions, was refined iteratively using the RTVE training data without the need of any transcriptions. The town hall plenaries used as a starting point are characterized by structured discourses with long speaking turns. Unsurprisingly, the system performed better on RTVE2018 data, which is closer to that style. However, the proposed approach proved to be effective, adapting to the new domain and providing better results at each iteration. If we focus, for example, on the TV show *Comando Actualidad* (CA), which is present in both datasets, we can see that the %WER obtained with the off-the-shelf system was 72.6%, 52.9% for the first self-training iteration and 36.27% for the last one. This suggest that the proposed strategy could be used to successfully adapt to new challenging domains, like the 2020 evaluation set, by leveraging on-domain unlabelled data and performing more self-training iterations.

6. Acknowledgements

The authors would like to thank the organizers of the Albayzin Challenge.

7. References

[1] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," 2015.

[2] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," 2005.

[3] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[4] H. Xu, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey, and S. Khudanpur, "A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5929–5933.

[5] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," 04 2018, pp. 5329–5333.

[6] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi," 09 2016, pp. 2751–2755.

[7] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," 09 2018, pp. 3743–3747.

[8] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH 2015 - 15th Annual Conference of the International Speech Communication Association*, September 2015.

[9] A. Stolcke, "Srilm — an extensible language modeling toolkit," *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, vol. 2, 07 2004.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

The Vicomtech Speech Transcription Systems for the Albayzín-RTVE 2020 Speech to Text Transcription Challenge

Aitor Álvarez, Haritz Arzelus, Iván G. Torre, Ander González-Docasal

Vicomtech Foundation, Basque Research and Technology Alliance (BRTA),
Mikeletegi 57, 20009 Donostia – San Sebastián (Spain)

[aalvarez,harzelus,igonzalez,agonzalezd]@vicomtech.org

Abstract

This paper describes the Vicomtech's submission to the Albayzín-RTVE 2020 Speech to Text Transcription Challenge, which calls for automatic speech transcription systems to be evaluated in realistic TV shows.

A total of 4 systems were built and presented to the evaluation challenge, considering the primary system along to three constrastive systems. These recognition engines are different versions, evolutions and configurations of two main architectures. The first architecture includes an hybrid DNN-HMM acoustic model, where factorized TDNN layers with and without initial CNN layers were trained to provide posterior probabilities to the HMM states. The language model for decoding correspond to modified Kneser-Ney smoothed 3-gram model, whilst a RNNLM model was used in some systems for re-scoring the initial lattices. The second architecture was based on the Quartznet architecture proposed by Nvidia with the aim of building smaller and ligther ASR models with SOTA-level accuracy. A modified Kneser-Ney smoothed 5-gram model was employed to re-score the initial hypothesis of this E2E model. The results obtained for each TV program in the final test set are also presented in addition to the hardware resources and computation time needed by each system to process the released evaluation data.

Index Terms: albayzín evaluations, speech recognition, deep learning, convolutional neural networks, recurrent neural networks.

1. Introduction

The Albayzín-RTVE 2020 Speech to Text Transcription Challenge calls for Automatic Speech Recognition (ASR) systems that are robust against realistic TV shows. Currently, it is a notable trend that aims to approach ASR technology to automate different applications such as subtitling or metadata generation for archive. Although most of this work is still performed manually or through semiautomatic methods (e.g. re-speaking), the current state of the art in speech recognition suggests that this technology can be exploitable autonomously without any post-edition task, mainly on contents with optimal audio quality and clean speech conditions. The use of Deep Learning algorithms in speech processing have made it possible to introduce this technology in such a complex scenario through the use of systems based on Deep Neural Networks (DNNs) or more recent architectures based on the End-To-End (E2E) principle.

During the last years, ASR systems have positively evolved at acoustic modeling with the integration of DNNs in combination with Hidden Markov Models (HMMs) to outperform traditional approaches [1]. More recently, new attempts have been focused on building E2E ASR architectures [2], which directly map the input speech signal to character sequences and therefore greatly simplify training, fine-tuning and inference [3, 4, 5, 6]. Additionally, deep Transformer or LSTM-RNN based language models have shown better performance than the traditional n-gran models specially during the re-scoring of the initial lattices [7].

Driven by the need to reduce the size and complexity of the ASR models, new architectures have recently arisen to make these models lighter, faster and more feasible to deploy on hardware with limited computation capabilities while maintaining the SOTA-level accuracy. In this sense, based on the Jasper architecture [8], Nvidia proposed Quartznet [9], a new E2E neural acoustic model composed of multiple blocks with residual connections in between. Each block consists of one or more modules with 1D time-channel separable convolutional layers, batch normalization, and ReLU layers. They reached near-SOTA error rates on the well-known LibriSpeech [10] and WSJ [11] datasets with models containing fewer than 20 million parameters, in contrast to other larger E2E architectures such as 5x3 Jasper (201 millions) [8] , Deep Speech 2 (38 millions) [2] or Wav2Vec2.0 (95 to 317 millions) [12].

Our systems were built following both DNN-HMM and Quartznet E2E architectures basis, in order to compare the performance of both systems trained with the same corpora as well as their feasibility to be deployed in different platforms, from high-performance servers to embedded systems like Nvidia's Jetson, Google's Coral or Intel's Movidious, among others. We presented a total of 4 ASR engines to the evaluation challenge; three systems based on DNN-HMM hybrid acoustic models, and one system constructed following the E2E Quartznet architecture.

The remainder of this paper is organised as follows: Section 2 describes the corpora used to train the systems; in Section 3 we describe the different speech transcription systems built for the challenge and Section 4 presents the results on the final evaluation test set, in addition to the number of resources and processing time needed per system to process the whole test set. Finally, Section 5 draws the main conclusions.

2. Corpus description

Since in this edition of the Albayzín-RTVE 2020 Speech to Text Transcription Challenge the *open training* condition was only considered, different corpora were used and mixed to train the acoustic and language models.

2.1. Acoustic corpus

The acoustic corpus was composed by annotated audio contents from 7 different datasets, summing up a total of 743 hours and 35 minutes. The following table 1 presents the final number of

10.21437/IberSPEECH.2021-22

hours containing only speech in each of the datasets.

Table 1: *Duration of the speech segments for each dataset*

dataset	duration
RTVE2018	112 h. 30 min.
SAVAS	160 h. 58 min.
IDAZLE	137 h. 8 min.
A la Carta	168 h. 29 min.
Common Voice	158 h. 9 min.
Albayzin	5 h. 33 min.
Multext	0 h. 47 min.
Total	743 h. 35 min.

The *RTVE2018* dataset [13] was released by RTVE and comprises a collection of TV shows drawn from diverse genres and broadcast by the public Spanish National Television (RTVE) from 2015 to 2018. This dataset was originally composed by 569 hours and 22 minutes of audio with a high portion of imperfect transcriptions and, thus, they could not be used as such for training. Therefore, a forced-alignment was applied in order to recover only the segments transcribed with a high literality, obtaining a total of 112 hours and 30 minutes of nearly correctly transcribed speech segments.

The *SAVAS* corpus [14] is composed of broadcast news contents in Spanish from 2011 to 2014 of the Basque Country's public broadcast corporation EiTB (Euskal Irrati Telebista), and includes annotated and transcribed audios in both clear (studio) and noisy (outside) conditions. The *IDAZLE* corpus is integrated by TV shows from the EiTB broadcaster as well, and it comprises a more varied and rich collection of programs of different genres and styles. TV shows are also the contents which compose the *A la Carta*[1] acoustic corpus, including 265 contents broadcasted between 2018 and 2019 by RTVE.

The *Common Voice* dataset [15] is a crowdsourcing project started by Mozilla to create a free and massively-multilingual speech corpus to train speech recognition systems. Finally, the well-known and clean *Albayzin* [16] and *Multext* [17] datasets were also included, mainly to favour the initial training steps and alignments of the systems.

2.2. Text corpus

Regarding text data, different sources were employed to obtain the enough language and domain coverage as close as possible to the contents of the challenge. The following Table 2 presents the number of words provided by each of the text corpus.

Table 2: *Description of the text corpus*

corpus	#words
Transcriptions	7,946,991
RTVE2018	56,628,710
A la Carta	106,716,060
Wikipedia	489,633,255
Total	660,925,016

A total of almost 661 million words were thus compiled and used to estimate the language models for decoding and rescoring purposes. The *Transcriptions* text corpus corresponded to the text transcriptions of the all audio contents used to train the acoustic models. The *RTVE2018* text corpus contains all the text transcriptions and re-spoken subtitles included within the RTVE2018 dataset, whilst the *A la Carta* corpus is integrated by subtitles taken from the "A la Carta" web portal, as a result of a collaboration between RTVE and Vicomtech. Finally, the *Wikipedia* corpus contained texts of the Wikipedia portal gathered in 2017 from Wikimedia[2].

3. Systems description

Two main architectures were employed to build the 4 systems presented to the Albayzín-RTVE 2020 Speech to Text Transcription Challenge: a hybrid DNN-HMM acoustic model for 3 of the systems, and Nvidia's Quartznet architecture for the final system.

3.1. DNN-HMM based systems

The DNN-HMM based systems were built through the *nnet3* DNN setup of the Kaldi recognition system [18], and using the so-called *chain* acoustic model based on optional Convolutional Neural Network (CNN) layers and a factorised time-delay neural network (TDNN-F) [19] which reduces the number of parameters of the network by factorising the weight matrix of each TDNN layer into the product of two low-rank matrices.

Two types of DNN-HMM acoustic models were constructed. The first architecture integrated a CNN-TDNN-F based network, with 6 CNN layers followed by 12 TDNN-F layers. Meanwhile, the second architecture integrated a TDNN-F model and consisted of 16 TDNN-F layers. In both systems, the internal cell-dimension of the TDNN-F layers was of 1536, with a bottleneck-dimension of 160 and a dropout schedule of '0,0@0.2,0.5@0.5,0'. The number of training epochs was set to 4, with a learning rate of 0.00015 and a minibatch size of 64. The input vector corresponded to a concatenation of 40 dimensional high-resolution MFCC coefficients, augmented through speed (using factors of 0.9, 1.0, and 1.1) [20] and volume (with a random factor between 0.125 and 2) [21] perturbation techniques, and the appended 100 dimensional iVectors.

The presented *primary* system was a CNN-TDNN-F based engine with a 3-gram LM for decoding and a 4-gram pruned RNNLM model used for lattice-rescoring following the work presented in [22]. The 3-gram LM was trained with texts coming from the *Transcriptions, RTVE2018* and *A la Carta* corpora presented in Table 2, and the 4-gram pruned RNNLM model was estimated adding the *Wikipedia* text corpus as well.

The first *contrastive* system included the same language models for decoding and rescoring as the *primary* system. The difference relied on the acoustic model, which corresponded to a TDNN-F based network without the initial CNN layers. Finally, the second *contrastive* system was a CNN-TDNN-F based system with a 3-gram LM for decoding and without applying any lattice-rescoring process.

3.2. Quartznet architecture based system

The E2E architecture, which corresponded to the submitted third *contrastive* system, was based on a 5x5 Quartznet system [9] which is completely based on 1D Time-Channel Separable Convolutional layers with residual connections. This design is based on the Jasper architecture [8] but with many modifications focused on considerably reducing the number of parameters and therefore, the computing resources needed.

[1] https://www.rtve.es/alacarta/

[2] https://dumps.wikimedia.org/

Initially there is a 1D Convolutional layer (kernel(k)= 33, output channels (c)= 256) processing the spectrogram input followed by five blocks with residual connections between blocks. Each block is composed of a module repeated five times. Each one of this modules are sequentially composed of (i) a k-sized depthwise convolutional layer, (ii) a pointwise convolution, (iii) a batch normalization layer, and (iv) a ReLU. The configuration of each Block is: B1 ($k = 33$, $c = 256$), B2 ($k = 39$, $c = 256$), B3 ($k = 51$, $c = 512$), B4 ($k = 63$, $c = 512$) and B5 ($k = 75$, $c = 512$). Finally there are three aditional convolutional layers: C1 ($k = 87$, $c = 512$), C2 ($k = 1$, $c = 1025$) and C3 ($k = 1$, $c = labels$). A Connectionist Temporal Classification (CTC) loss function is used for measuring prediction errors and Novograd optimizer with betas 0.8 and 0.5 is used for training with 100 epochs cosine annealing learning rate policy. Initial learning rate was set to 0.015, and minimum to 10^{-5}, weight decay was 10^{-3} and training dataset was computed on three GPUs with batch size of 40 each and mixed precision. Our resulting $5x5$ Quartznet network configuration contains $6,7$ million parameters.

Additionally, a 5-gram external language model, trained with the *Transcriptions*, *RTVE2018* and *A la Carta* corpora, was used during inference for rescoring the initial hypothesis by using Beam Search CTC Decoder with a beam-width of 1000, $\alpha = 1.2$ and $\beta = 0$. It is worth mentioning that in the previous DNN-HMM based systems, we could not use a 5-gram as LM for decoding due to the lack of memory resources to generate such a large graph with Kaldi.

4. Results and resources

In the following Table 3, the total WER values are presented for each submitted system over all the TV programs in the test set.

Table 3: *Total WER results per system on the whole Albayzin-RTVE 2020 testset*

type	***system***	***tWER***
P	Vicomtech_p-CNN_TDNN_Rescoring	**19.27**
C1	Vicomtech_c1-TDNN_Rescoring	19.98
C2	Vicomtech_c2-CNN_TDNN	19.83
C3	Vicomtech_c3-Quartznet	28.42

As it can be appreciated in Table 3, the *primary* system was correctly selected by the participants since it was the system with the best performance. Likewise, as expected, the Quartznet based experimental system achieved the worse results, even though the quality reached seems promising considering the resources and computing time needed for inference, in addition to the lightness of its E2E model. The robustness of the this E2E model could be improved by adding more training data.

It is also worth noting how the first *contrastive* system performs worst than the second *contrastive* even though the initial results were rescored by a RNNLM model in the former. It seems that, in this case, the acoustic model, which integrated CNN convolutional layers, helped more the ASR engines than rescoring the initial lattices at language model level. It could make sense for this test set, since most of the evaluation contents included spontaneous speech and our rescoring language models were trained with formal language gathered mostly from the *Wikipedia* encyclopedia.

In Table 4, the total WER results obtained by the *primary* system for each TV program in the Albayzin-RTVE 2020 test set are described. The behaviour of the *primary* system regarding the content profiles is similar as expected. In those programs with clean speech, the WER decreases significantly compared to other programs which included adverse acoustic conditions, overlapping or spontaneous speech. More specifically, in TV shows such as *Aquí la Tierra*, *Los desayunos de TV*, *Millenium* and *Versión Española*, with controlled acoustic conditions (*studio*) and long segments with dictate and well-structured speech, the word error rates are below the 20% border in all cases. In contrast, in more complicated contents to process automatically like *Cómo nos Reíamos*, *Imprescindibles* or *Wake-up*, which include many segments with spontaneous and acted speech, acoustically adverse conditions and overlapping, the results degrade appreciably.

Table 4: *Total WER on each test TV program of the Albayzin-RTVE 2020 testset by the primary system*

TV Programs	***tWER***
Aquí la Tierra	16.48
Boca Norte	37.94
Bajo la Red	33.31
Comando Actualidad	24.68
Como nos Reíamos	48.53
Ese Programa del que Usted me Habla	25.67
Imprescindibles (live recordings)	34.45
Los desayunos de TV	10.11
Mercado central	17.83
Millennium	15.98
Never Films Mira Ya	24.21
Si Fueras Tu	29.31
Vaya Crack	19.96
Versión Española	18.10
Wake-Up	33.96
Global	**19.27**

Nevertheless, the global total WER reached 19.27%, an interesting mark considering the difficulty of the test set and that the same engines and models were used within each ASR system to transcribe such different contents from each other in terms of domains, acoustic conditions and speech type. It could have been interesting to check if using different language models fine-tuned to specific domains (debates, news, comedy shows, etc.) and applying them to the corresponding type of contents, the results would have improved.

4.1. Processing time and resources

The decoding processes of the 4 transcription systems were performed on an Intel Xeon CPU E5-2683v4 2.10 GHz 7xGPU server with 256GB DDR4 2400MHz RAM memory. The GPU used for decoding corresponds to an NVIDIA Geforce GTX 1080 Ti 11GB graphics acceleration card.

The following Table 5 presents the processing time and computational resources needed by each submitted system for the decoding of the released *evaluation set* of 55.9 hours of audios. It should be noted that the DNN-HMM based systems were decoded using one CPU core, whilst the Quartzent E2E systems took advantage of a GPU card. In terms of Real-Time Factor (RTF), while the Kaldi-based *primary* system achieved a 0.98 of RTF, the Quartznet based engine reached 0.13 of RTF to process the whole evaluation contents.

Table 5: *Processing time and computational resources needed by each submitted system*

system	*RAM (GB)*	*CPU cores*	*GPU (GB)*	*Time*
p-CNN_TDNN_Rescoring	6.7	1	-	55h
c1-TDNN_Rescoring	6.7	1	-	49h
c2-CNN_TDNN	5.9	1	-	39h
c3-Quartznet	6	1	9.9	7.5h

5. Conclusions

Vicomtech submitted 4 transcription systems; three systems based on more traditional Kaldi's DNN-HMM based engines, using a 3-gram LM for decoding and a RNNLM for lattices rescoring, and a more experimental last system inspired on an optimization of the Nvidia's Quartznet E2E architecture, which aims to be deployed in embedded systems with remarkably accurate results.

As expected, the error rates of the three former systems were notably lower comparing to the E2E based model. However, the participants were motivated to evaluate this novel architecture, designed to be lighter than traditional ASR engines, in order to check their robustness in the same training and evaluation conditions. Nowadays, as larger neural networks with more layers and parameters are built, reducing their complexity and computational cost has becoming critical, specially in real-time applications and scenarios. In addition, the evolution of embedded systems with high computational capacities, triggered great opportunities for researchers to face fundamental challenges in deploying deep learning systems for portable devices with limited resources.

6. References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[3] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14. JMLR.org, 2014, pp. II–1764–II–1772.

[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4960–4964.

[5] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[6] L. Lu, X. Zhang, and S. Renals, "On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5060–5064, 2016.

[7] Y. Wang, A. Mohamed, D. Le, C. Liu, A. Xiao, J. Mahadeokar, H. Huang, A. Tjandra, X. Zhang, F. Zhang *et al.*, "Transformer-based acoustic modeling for hybrid speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6874–6878.

[8] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, "Jasper: An end-to-end convolutional neural acoustic model," *arXiv preprint arXiv:1904.03288*, 2019.

[9] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6124–6128.

[10] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[11] D. B. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[12] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[13] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. de Prada, "Albayzin 2018 evaluation: the iberspeech-rtve challenge on speech technologies for spanish broadcast media," *Applied Sciences*, vol. 9, no. 24, p. 5412, 2019.

[14] A. del Pozo, C. Aliprandi, A. Álvarez, C. Mendes, J. P. Neto, S. Paulo, N. Piccinini, and M. Raffaelli, "Savas: Collecting, annotating and sharing audiovisual language resources for automatic subtitling." in *LREC*, 2014, pp. 432–436.

[15] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[16] F. Casacuberta, R. Garcia, J. Llisterri, C. Nadeu, J. Pardo, and A. Rubio, "Development of spanish corpora for speech research (albayzin)," in *Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assesment Methods, Chiavari, Italy*, 1991, pp. 26–28.

[17] E. Campione and J. Véronis, "A multilingual prosodic database," in *Fifth International Conference on Spoken Language Processing*, 1998.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[19] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Interspeech*, 2018, pp. 3743–3747.

[20] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[21] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[22] H. Xu, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey, and S. Khudanpur, "A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5929–5933.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Sigma-UPM ASR Systems for the IberSpeech-RTVE 2020 Speech-to-Text Transcription Challenge

Juan M. Perero-Codosero[1,2], *Fernando M. Espinoza-Cuadros*[1,2], *Luis A. Hernández-Gómez*[2]

[1]Sigma Technologies S.L.U., Madrid, Spain
[2]GAPS Signal Processing Applications Group, Universidad Politécnica de Madrid, Spain
{jmperero,fmespinoza}@sigma-ai.com, luisalfonso.hernandez@upm.es

Abstract

This paper describes the Sigma-UPM Automatic Speech Recognition (ASR) systems submitted to IberSpeech-RTVE 2020 Speech-to-Text Transcription Challenge. Deep Neural Networks (DNNs) are becoming the most promising technology for ASR at present. Since last few years, traditional hybrid models are being evaluated and compared to other end-to-end ASR systems in terms of accuracy and efficiency. In this challenge, we contribute with two different approaches: a primary hybrid ASR system based on DNN-HMM and two contrastive state-of-the-art end-to-end ASR systems, based on lattice-free maximum mutual information (LF-MMI). Our analysis of the results from the last edition led us to conclude that some adaptation should be accomplished to improve the performance of the systems. In particular, data augmentation techniques and Domain Adversarial Training (DAT) have been applied to the aforementioned approaches. Multi-condition data augmentation applied to our hybrid DNN-HMM models has demonstrated WER improvements in noisy scenarios (about 10% relative). In contrast, results obtained using an end-to-end Pychain-based ASR system are far from our expectations. Nevertheless, we found that when including DAT techniques a relative WER improvement of 2.87% was obtained as compared to the Pychain-based system.

Index Terms: TV shows Speech-to-Text transcription, ASR systems, Hybrid DNN-HMM, end-to-end Deep Learning, Domain Adversarial Training

1. Introduction

State-of-the-art ASR approaches are mainly based on DNNs. Particularly, in traditional hybrid systems acoustic models use the Hidden Markov Model (HMM) state probabilities to train a DNN. These DNN-HMM acoustic models (AM) are combined with other models: pronunciation (PM) and language models (LM). According to some research studies [1] these hybrid-models perform better in many scenarios with small amount of training data.

Nevertheless, hybrid ASR systems have some limitations such as a high complexity associated to the training process of DNN-HMM models. They require phoneme alignments for frame-wise cross entropy and a sophisticated beam search decoder [2]. Furthermore, they usually require strong context-dependency trees to train *chain models* [3].

Trying to overcome these limitations, end-to-end ASR systems have been emerging in the last years. Several approaches appeared such as Connectionist Temporal Classification (CTC) [4], Recurrent Neural Network Transducer (RNN-T) [5] and sequence-to-sequence attention-based encoder-decoder [6, 7]. This trend presents an easy-to-use and easy-to-update pipeline. First, training process has not several stages, only a single model is required instead of the traditional ones (e.g., AM, PM, LM). Second, the continuous advances in deep learning-based technologies have allowed the quick development of powerful open-source libraries for machine learning, such as PyTorch [8] or TensorFlow [9], among others.

The promising results reported for many end-to-end ASR systems depend on the scenarios as well as the availability of datasets. There is an evident gap between these end-to-end systems and hybrid models. Nevertheless, recent developments focused on reducing this gap have achieved good results, as is the case of Pychain [10]. In Pychain, the end-to-end LF-MMI criterion, which is the state-of-the-art for hybrid models in Kaldi [3], is implemented by combining a single stage training and a fully parallelization under PyTorch framework.

Besides that, in order to improve the performance of end-to-end ASR systems, a variety of techniques commonly applied in Deep Learning have been introduced. Data augmentation techniques [11, 12] have been developed to increase the quantity of training data following some criteria to improve model robustness. Thus, a variety of scenarios can be simulated trying to cover more challenging acoustic conditions in a cost-effective way. Other recent deep learning-based technique, namely Domain Adversarial Training (DAT) [13] has been applied to improve ASR performance by learning features invariant to different conditions, such as acoustic variabilities [14, 15], accented speech [16], and inter-speaker feature variability [17].

In this paper, our aim is to contribute to the evaluation of both hybrid and end-to-end ASR systems under the conditions of the IberSpeech-RTVE 2020 Speech-to-Text Transcription Challenge [18]. For this purpose, we firstly report the use of data augmentation techniques to improve our Kaldi-based hybrid ASR system presented in the previous IberSpeech-RTVE edition [19]. Then we evaluate a baseline Pychain system and an improved version of it including DAT. These evaluations on IberSpeech-RTVE Challenge allow us to compare the performance of our systems in a variety of TV shows and broadcast news in specific conditions, noisy environments or challenging scenarios.

2. Architectures

2.1. Primary system: Hybrid ASR

This system is based on the Sigma ASR [19] submitted to the Albayzin-RTVE 2018 Speech-to-Text Challenge [20], where it was in the top-2 of the ranking for both closed and open condition evaluations.

This hybrid ASR system was built using Kaldi Toolkit [1]. The architecture consists of the classical sequence of three modules: acoustic model, pronunciation model and language model.

10.21437/IberSPEECH.2021-23

The acoustic model is based on Deep Neural Networks and Hidden Markov Models (DNN-HMM) following the so-called *chain models* [3].

Our main conclusion from Albayzin-RTVE 2018 results was the need for more robust DNN training looking for accuracy improvements required in the more challenging scenarios (street interviews, game shows, risky sports documentaries...). Environmental robustness of acoustic models has been significantly improved by using multi-condition training data. However, data collection processes are very costly compared to the artificial generation of new training data, which has become a common alternative [11].

Thus, in our current contribution we extended the amount of training data through data augmentation techniques. In particular, we added reverberation to the available training speech data, following the approach presented in [21]. Depending on the expected scenarios and distances, different room impulse responses (RIRs) can be used. The authors sample the room parameters and receiver position in the room and then randomly generate a number of RIRs according to different speaker positions. Basically, three sets of simulated RIRs have been applied: small room (1-10 m), medium room (10-30 m) and large room (30-50 m). The real computation is carried out at the feature extraction level where the original data is mixed with its reverberated copies. The result is a 2-fold training set.

This data augmentation technique has been added to other data augmentation techniques already used in the recipe followed in our previous system, such as volume and speed perturbations [11, 12].

2.2. Contrastive system 1: end-to-end LF-MMI ASR

Aiming to explore the new state-of-the-art end-to-end ASR systems, we have evaluated an alternative to the developed Kaldi-based hybrid ASR system. That is the new end-to-end ASR lattice-free maximum mutual information (LF-MMI) approach [3], which is also used in Kaldi's *chain models*. Thus, we found it interesting to perform a reliable comparison of these two systems based on LF-MMI under RTVE IberSpeech 2020 Challenge scenarios.

This first end-to-end contrastive system is based on PyChain [10], a powerful PyTorch-based implementation which is intended to have an easy-to-use pipeline in which data preparation and final decoding are carried out in Kaldi for efficiency, while data loading and network training are made in PyTorch [8]. It should be noted that it is not necessary any alignment, i.e., HMM-GMM training stage is not required, unlike other systems [22, 23].

Data preparation consists in both feature extraction (40-dimensional MFCC) and the generation of the numerator/denominator graphs stored as Finite State Transducers (FSTs). By following the Kaldi's method for LF-MMI, HMM graphs are used as supervision. The probability distribution function (pdf) is used to estimate the likelihood of an HMM emission [1]. In this case, the network output and the occupation probability are computed from a pdf-index (pdf-id) instead of an HMM state. More details are presented in [10].

Once the data is loaded, the Pytorch model tries to simulate a time-delay neural network (TDNN) [24] by including 1D dilated convolution in addition to batch normalization, ReLU and dropout. This sequence is stacked in this order up to six layers with residual connections. At the end of the sequence, a fully connected layer is added (as described in [10]). From now on this system will be called Pychain-based baseline system.

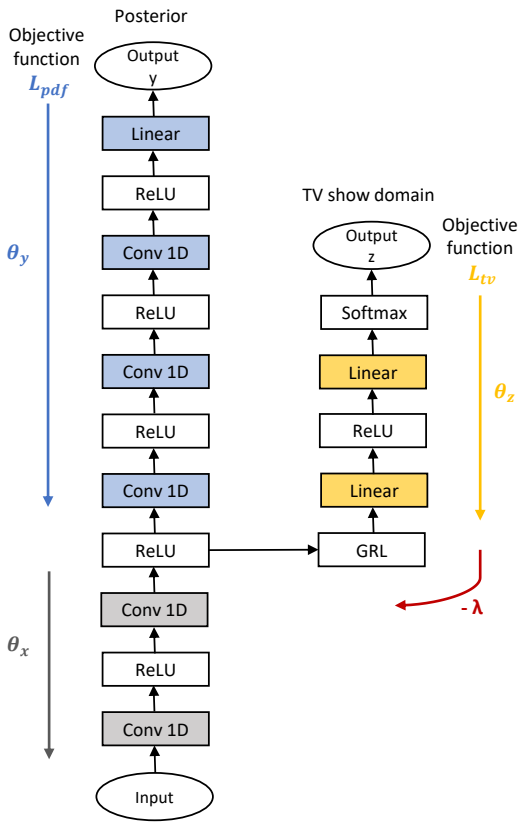


Figure 1: *Architecture of the contrastive system 2. An adversarial branch (TV show classifier) is added to the second layer of the main Pychain architecture (posterior classifier).*

2.3. Contrastive system 2: Domain Adversarial Training

Different acoustic conditions on TV shows can have a negative impact on the Pychain-based baseline performance. In order to reduce this effect, we explore the integration of Domain Adversarial Training (DAT) [13] trying to improve the Pychain-based baseline system. More specifically, in this approach we aim to generate acoustic representations invariant to the domain of the TV show characteristics by using a Domain Adversarial Neural Network (DANN).

For this adversarial architecture, a training dataset denoted as $\{x_i, y_i, z_i\}_N^{i=1}$ is composed of x_i, which are the acoustic features, and y_i, z_i which are the posteriors of the senones and the type of TV show, respectively.

Differently from the Pychain-based baseline system training, in which the acoustic representation is trained so as to minimize the LF-MMI loss function, in DAT the acoustic representations are learned adversarially to the secondary task (i.e., TV show classification). In this way, the domain-dependent information is supressed from the representation as it is irrelevant for the primary task (i.e., posterior classifier).

As it can be seen in Figure 1, the parameters of our adversarial architecture consist of three parts, $\theta = \{\theta_x, \theta_y, \theta_z\}$. Where θ_x denotes the parameters of the first layers of the TDNN used as feature extractor, and θ_y and θ_z denote the parameters of the pdf posteriors and the TV show classifier sub-networks respectively.

A gradient reversal layer (GRL) [13] is placed between the

feature extractor and the TV show classifier. In the forward propagation, GRL keeps the input unchanged and reverses the gradient by multiplying it by a negative coefficient during the backpropagation.

According to [13], for this adversarial training, the objective functions for the pdf classifier L_{pdf} and TV show classifier L_{tv} are defined as:

$$L_{pdf}(\theta_x, \theta_y) = -\sum_{i=i}^{N} \log P(y_i|x_i; \theta_x, \theta_y) \quad (1)$$

$$L_{tv}(\theta_x, \theta_z) = -\sum_{i=i}^{N} \log P(z_i|x_i; \theta_x, \theta_z) \quad (2)$$

DNN acoustic model and the adversarial branch are jointly trained to optimize the following:

$$\min_{\theta_x, \theta_y} \max_{\theta_z} L_{pdf}(\theta_x, \theta_y) - \lambda L_{tv}(\theta_x, \theta_z), \quad (3)$$

where λ is a trade-off parameter between the pdf classification loss L_{pdf} corresponding to the LF-MMI loss, and the TV show classification loss L_{tv}, whose goal is to make deep acoustic features invariant to the domain of the TV show characteristics.

3. Experimental setup

3.1. Datasets

The proposed ASR systems have been evaluated under the Albayzin-RTVE 2020 Challenge conditions. RTVE2020 Database [25] has been provided to all the participants. It is an extension of RTVE2018 Database which contains a collection of TV shows and broadcast news from 2015 to 2019.

RTVE2018 training partition was prepared under a manual supervision process [19] in order to have reliable transcriptions aligned with the speech signal. Two datasets are used for the system evaluation: RTVE_train350 (350 hours of train set) and RTVE_train100 (a RTVE_train350 subset of 100 hours). Validation datasets are a 20% of training data. For testing purposes, several datasets corresponding to 1 hour of duration each were built from RTVE_dev1 and RTVE_dev2 partitions. Moreover, the whole RTVE2020 database was used as test partition for this challenge. It is composed of more than 70 hours of audio and it has been used to present results after the submission.

Additional datasets were added to train the system in open training condition scenario, as used in [19]: VESLIM [26] (103 hours of Spanish clean voice) and OWNMEDIA (162 hours of TV contents, interviews and lectures).

Data augmentation techniques related to hybrid ASR system are carried out by means of the reverberation database[1], which has been described in Section 2.1.

3.2. Training setup

The acoustic model of the hybrid ASR system was trained using RTVE_train350, VESLIM and OWNMEDIA databases, following the SWBD Kaldi recipe for *chain models*. Some modifications were included following the ASpIRE recipe for multi-condition tasks. The rest of the setup is the same as in [19].

End-to-end LF-MMI models (both contrastive systems) were trained by using only RTVE_train100. This relatively small amount of data allows a light training process to test the system performance. We use Pychain-example[2] as a reference by adding some changes in terms of data loading in order to use more than one GPU in parallel.

Data preparation is carried out in Kaldi. In order to convert input features into PyTorch tensors, we use kaldi_io[3] as suggested in [10].

For the adversarial training, note that that the number of the pdf posteriors is $y_i = 62$, and the number of TV shows is $z_i = 13$ because of the different TV shows that RTVE_train100 partition contains. To reduce the bias effect into system training due to unbalanced classes in TV shows, the data was previously merged according to their acoustic characteristics. As a result, four new groups were defined: 0) interviews, 1) TV-game shows, 2) documentaries, and 3) live TV shows. Thus, the labels of the training data for the adversarial branch (i.e., the TV show classifier sub-network) are defined as $z_i = \{0, 1, 2, 3\}$.

In the adversarial architecture, the second hidden layer of the TDNN is used as input to the adversarial branch, which consists of a dense layer of size 384 and ReLU activation function followed by a *softmax* output layer, whose output dimension corresponds to the number of TV shows (i.e., 4). Cross-entropy loss function was used on the adversarial training. To select the optimal trade-off parameter λ, several values were tested. The best performance was achieved for $\lambda = 0.041$.

In addition, all the systems have been evaluated with the same N-gram LM. As described in [19], several corpora were used: subtitles provided in RTVE2018 Database, supervised transcriptions, news between 2015 and 2018, interviews and file captions. A selected lexicon containing 120k words was extracted to train LMs. A 3-gram LM was trained for decoding stage in both hybrid ASR and Pychain-based systems. Instead, a 4-gram LM was only used for the Pychain rescoring stage.

3.3. Resources

Several computation resources have been required to carry out this work. A server with 2 Xeon E5-2630V4, 2.2 GHz, 10C/20 TH and 3 GPUs Nvidia GTX 1080 Ti. Regarding the hybrid ASR system, GPU calculation is necessary for the DNN stage and only CPU mode is used for the HMM stage and final decoding. In case of the Pychain-based systems, only 2 GPUs were used for both training and decoding stages.

4. Results

4.1. Hybrid ASR

The proposal of applying data augmentation techniques to improve the hybrid ASR performance has been fulfilled. The addition of reverberation to our whole training dataset (over 600 hours of speech) has improved the performance in most of the scenarios represented by every TV show set. As shown in Table I, the model applied to CA (*Comando Actualidad*) dataset achieved a relative improvement around 10%, as compared to the baseline system. This might be due to the trained model having learned to model these speech artifacts which can appear in challenging scenarios described in Section 2.1. However, the improvements in the rest of TV shows are not so remarkable (e.g., 20H or LM) because contents are related to daily news with more favorable acoustic conditions.

As we mentioned in [19], the reference master of transcriptions was not reviewed. As usual, we have evaluated the possible impact of transcription errors by means of a new test using an external dataset. It consists of 3.5 hours of TV news

[1] http://www.openslr.org/28/

[2] https://github.com/YiwenShaoStephen/pychain_example

[3] https://github.com/vesis84/kaldi-io-for-python

Table I: *WER(%) on the different datasets for hybrid and end-to-end ASR systems. The duration of each dataset is an hour.*

	20H_dev1	**AP_dev1**	**CA_dev1**	**LM_dev1**	**Mill_dev1**	**LN24H_dev1**
Hybrid ASR						
Kaldi-based baseline [19]	14.88	20.94	49.55	21.44	17.01	24.13
Reverb. data augmentation	**14.76**	21.00	**44.69**	**21.03**	**16.42**	**23.62**
Kaldi-based baseline (RTVE_train100)	16.09	22.32	51.23	23.02	17.70	25.53
End-to-end LF-MMI ASR						
Pychain-based baseline	23.66	33.31	59.34	29.95	35.09	25.08
Domain Adversarial Training	**23.53**	**32.99**	**59.25**	**29.91**	**34.67**	25.16

Table II: *WER(%) on the RTVE2020 test partition for the all the systems. Results were obtained after the submission.*

	RTVE2020_test
Hybrid ASR	
Kaldi-based baseline [19]	31.01
Reverb. data augmentation	27.68
End-to-end LF-MMI ASR	
Pychain-based baseline	40.90
Domain Adversarial Training	42.89

Table III: *Real-time factor (RT) for the different stages carried out in the Pychain-based baseline according to the different datasets.*

Datasets	**Decoding**	**LM rescoring**
20H_dev1	0.033	0.115
AP_dev1	0.035	0.175
CA_dev1	0.225	1.976
LM_dev1	0.092	0.450
Mill_dev1	0.082	0.442
LN24H_dev1	0.383	0.148

broadcasts (similar to 20H). Applying the reverb-trained models of our primary system, we reduced the WER from 8.51% to 7.96%, being our new best results achieved so far.

After the submission, we also evaluated this system on RTVE2020 test partition which contains mostly challenging TV contents. Table II shows data augmentation maintains the WER improvement around 10% relative.

4.2. End-to-end LF-MMI ASR

Our Pychain-based baseline has a good performance in relation to the number of parameter and the easier training process as compared to other end-to-end frameworks. The WER achieved for standard TV news (e.g., 20H, LN24H) are between 23% and 26%, as shown in Table I. These results are within the expected range where commercial ASR systems operate.

In order to compare both hybrid and end-to-end ASR systems, we also trained a hybrid model by using only RTVE_train100 partition. In this case, multi-condition data augmentation has not been applied. Pychain-based system is still far from Kaldi-based hybrid system, with a WER increase of 17% in the worst-case scenario. This gap could be reduced with the application of some data augmentation techniques (e.g., speed or volume perturbation). Despite the fact that augmented data causes a slightly improvement on WER for Pychain system [10], TV shows can contain some acoustic characteristics which could be better modeled by some audio perturbations.

Regarding time consumption, Pychain carries out two stages: a decoding stage and a 4-gram rescoring stage. Table III shows time requirements depends on the characteristics of the test partitions. Less time is required for good acoustic conditions and less challenging scenarios. It makes sense as far as the confusion of the graph model is less complex to transcribe accurately. This behaviour is the same when applying DAT.

On the other hand, we evaluated the effect of learning acoustic representations invariant to TV shows domain. After applying DAT, results in Table I show improvements, in terms of WER, up to 2.87% as compared to the Pychain-based baseline. We are aware that those results are far from the primary hybrid ASR system based on DNN-HMM. Nevertheless, results in Table I give us an insight on how the use of DAT into the end-to-end LF-MMI model can improve its performance in most of the scenarios. It seems that DAT is able to generate deep acoustic features invariant to different TV shows with different acoustic conditions without the need of data augmentation techniques. In addition, Table II shows that applying DAT does not reduce the WER on RTVE2020 test partition. The main reason is DAT alleviates labeled domain conditions in the training dataset. Thus, invariant features were trained without regarding these unseen external factors.

The output transcriptions for this system were not submitted on time due to the required computational demand.

5. Conclusions

In this paper, we have developed both hybrid and end-to-end ASR approaches applying some techniques focused on improving the performance of Text-to-Speech task. Hybrid DNN-HMM models can be adapted to the TV show domain by means of multi-condition data augmentation. The addition of reverberated data to the training data decreases WER significantly (10% relative). A WER of 7.96% is achieved in better conditions. Moreover, we have demonstrated that the lack of data augmentation techniques could be the main reason of the gap between Kaldi hybrid system and Pychain-based system. However, other easy-to-apply techniques, such as DAT, can overcome this gap yielding improvements in end-to-end ASR systems. Acoustic features invariant to the TV show domain are learned by the model achieving a WER improvement of 2.87%. As future work, we believe that adding perturbations to the training data or exploring speech enhancement techniques could help to close the performance gap between Kaldi hybrid system and Pychain-based system. In addition, unsupervised machine learning methods or automatic perceptual speech quality methods could contribute to a more accurate TV shows classification prior to DAT.

6. References

[1] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[2] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," *arXiv preprint arXiv:1805.03294*, 2018.

[3] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in *Interspeech*, 2016, pp. 2751–2755.

[4] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[5] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[6] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[7] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[8] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.

[9] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

[10] Y. Shao, Y. Wang, D. Povey, and S. Khudanpur, "Pychain: A fully parallelized pytorch implementation of lf-mmi for end-to-end asr," *arXiv preprint arXiv:2005.09824*, 2020.

[11] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms." in *ASRU*, 2015, pp. 539–546.

[12] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[13] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[14] D. Serdyuk, K. Audhkhasi, P. Brakel, B. Ramabhadran, S. Thomas, and Y. Bengio, "Invariant representations for noisy speech recognition," *arXiv preprint arXiv:1612.01928*, 2016.

[15] Y. Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition." in *Interspeech*. San Francisco, CA, USA, 2016, pp. 2369–2372.

[16] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie, "Domain adversarial training for accented speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4854–4858.

[17] Z. Meng, J. Li, Z. Chen, Y. Zhao, V. Mazalov, Y. Gang, and B.-H. Juang, "Speaker-invariant training via adversarial learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5969–5973.

[18] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. de Prada, "Albayzin evaluation: Iberspeech-rtve 2020 speech to text transcription challenge," http://catedrartve.unizar.es/reto2020/EvalPlan-S2T-2020-v1.pdf, 2020, [Online].

[19] J. M. Perero-Codosero, J. Antón-Martín, D. T. Merino, E. L. Gonzalo, and L. A. H. Gómez, "Exploring open-source deep learning asr for speech-to-text tv program transcription." in *IberSPEECH*, 2018, pp. 262–266.

[20] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. de Prada, "Albayzin 2018 evaluation: the iberspeech-rtve challenge on speech technologies for spanish broadcast media," *Applied Sciences*, vol. 9, no. 24, p. 5412, 2019.

[21] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[22] M. Ravanelli, T. Parcollet, and Y. Bengio, "The pytorch-kaldi speech recognition toolkit," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6465–6469.

[23] D. Can, V. R. Martinez, P. Papadopoulos, and S. S. Narayanan, "Pykaldi: A python wrapper for kaldi," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5889–5893.

[24] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[25] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. de Prada, "Rtve2020 database description," http://catedrartve.unizar.es/reto2020/RTVE2020DB.pdf, 2020, [Online].

[26] D. T. Toledano, L. A. H. Gómez, and L. V. Grande, "Automatic phonetic segmentation," *IEEE transactions on speech and audio processing*, vol. 11, no. 6, pp. 617–625, 2003.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

BCN2BRNO: ASR System Fusion for Albayzin 2020 Speech to Text Challenge

Martin Kocour[1], *Guillermo Cámbara*[2,3], *Jordi Luque*[2], *David Bonet*[2], *Mireia Farrús*[4], *Martin Karafiát*[1], *Karel Veselý*[1] *and Jan "Honza" Černocký*[1]

[1]Brno University of Technology, Speech@FIT, Czechia
[2]Telefónica I+D, Research
[3]Universitat Pompeu Fabra
[4]Universitat de Barcelona

ikocour@fit.vutbr.cz

Abstract

This paper describes the joint effort of BUT and Telefónica Research on the development of Automatic Speech Recognition systems for the Albayzin 2020 Challenge. We compare approaches based on either hybrid or end-to-end models. In hybrid modelling, we explore the impact of a SpecAugment layer on performance. For end-to-end modelling, we used a convolutional neural network with gated linear units (GLUs). The performance of such model is also evaluated with an additional n-gram language model to improve word error rates. We further inspect source separation methods to extract speech from noisy environments (i.e. TV shows). More precisely, we assess the effect of using a neural-based music separator named Demucs. A fusion of our best systems achieved 23.33 % WER in official Albayzin 2020 evaluations. Aside from techniques used in our final submitted systems, we also describe our efforts in retrieving high-quality transcripts for training.

Index Terms: fusion, end-to-end model, hybrid model, semi-supervised, automatic speech recognition, convolutional neural network.

1. Introduction

This paper describes the BCN2BRNO's team Automatic Speech Recognition (ASR) system for the IberSPEECH-RTVE 2020 Speech to Text Transcription Challenge, a joint collaboration between Speech@FIT research group, Telefónica Research (TID) and Universitat Pompeu Fabra (UPF). Our goal is to develop two distinct ASR systems, one based on a hybrid model [1] and the other one on an end-to-end approach [2], and complement each other through a joint fusion.

We submitted one primary system and one contrastive system. The primary system – Fusion B – is a word-level ROVER fusion of hybrid ASR models and end-to-end models. It achieved 23.33 % WER on official evaluation dataset. However, the same result was accomplished by the contrastive system – Fusion A–, a fusion which comprises only hybrid ASR models. In this paper we describe both ASR systems and a post-evaluation analysis and experiments that lead to a better performance of the primary fusion. We also discuss the effect of speech enhancement techniques such as background music removal or speech denoising.

2. Data

The Albayzin 2020 challenge comes with two databases: *RTVE2018* and *RTVE2020*. RTVE2018 is the main source of training and development data, while the RTVE2020 database is used for the final evaluation of submitted systems. RTVE2018 database [3, 4] comprises 15 different TV programs broadcast by the Spanish public television Radiotelevisión Española (RTVE). The programs contain a great variety of speech scenarios from read speech to spontaneous speech, live broadcast or political debates. The database consists of 569 hours of audio data, from which 468 hours are provided with subtitles (train set), and 109 hours are human-revised (dev1, dev2 and test sets). Both hybrid and end-to-end models utilize dev1 and train sets for training, while dev2 and test sets serve as validation datasets. RTVE2020 database [5] consists of 70 hours of manually annotated TV shows broadcast by the RTVE.

In addition, three Linguistic Data Consortium corpora were used for training the language model in the hybrid ASR system: *Fisher Spanish Speech* [6], *CALLHOME Spanish Speech* [7] and *Spanish Gigaword Third Edition* [8]. The audio recordings from CALLHOME and Fisher Spanish Speech corpus were used only for training tiny ASR model for transcript retrieval.

The end-to-end model is trained on Fisher Spanish Speech, Mozilla's Common Voice Spanish corpus and Telefónica's Call Center in-house data (23 hours). The Mozilla's Common Voice Spanish [9] corpus is an open-source dataset that consists of recordings from volunteer contributors pronouncing scripted sentences, recorded at 48kHz rate. The sentences come from original contributor donations and public domain movie scripts. The version of Common Voice corpus used for this work is 5.1, which has 521 hours of recorded speech. However, we have kept only speech validated by the contributors, i.e. 290 hours.

2.1. Transcript retrieval

The training data from RTVE2018 database includes many hours of subtitled speech. However, the captions contain several errors. In most cases, captions are shifted by a few seconds, so a segment with correct transcript corresponds to a different portion of audio. This phenomenon also occurs in human-revised development and test sets. Another problem with subtitled speech is "partly-said" captions. This issue involves misspelled and unspoken words of the transcription.

Since the training procedure of the hybrid ASR system is quite error-prone in case of misaligned labels, we decided to apply a transcript retrieval technique developed by Manohar, et al. [10]: the closed-captions related to the same audio, i.e., the whole TV show, are first concatenated according to the original timeline. This creates a small text corpus containing a few hundreds of words. The text corpus is used for training a biased N-gram language model (LM) with $N = 7$, so the model is biased only on the currently processed captions. During decoding, the weight of the acoustic model (AM) is significantly smaller than

10.21437/IberSPEECH.2021-24

the weight of LM, because we believe that the captions should occur in hypotheses. Then, the "winning" path is retrieved from the hypothesis lattice as the path that has a minimum edit cost w.r.t. the original transcript. Finally, the retrieved transcripts are segmented using the CTMs obtained from the oracle alignment (previous step). The segments, which do not correspondent to original transcripts, are discarded in this step. More details can be found in [11, 10].

Table 1: *2-pass transcript retrieval.*

Cleaning	Train	Dev1	Dev2	Test
Original	468	60.6	15.2	36.8
1-pass	99.4	21	7.5	-
2-pass	234.2	55.1	14.3	33.7
Recovered	50 %	91 %	94 %	92 %

The transcript retrieval technique is applied twice. First, we train an initial ASR system on out-of-domain data, e.g., Fisher and CALLHOME. This system is used to obtain hypothesis in the 1st pass of transcript retrieval. Then, a new system is trained from scratch on clean data from 1st pass and the whole process of transcript retrieval is repeated again. Table 1 shows how this 2-pass cleaning leads to the recovery of almost all the manually annotated development data and half of the subtitled training data.

Figure 1: *Amount of cleaned audio per TV-show, in hours.*

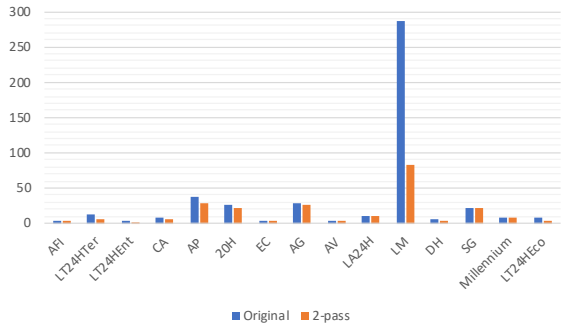


Figure 1 depicts how many hours have been recovered in individual TV programs. It also shows how data is distributed in the database. Most speech comes from La-Mañana (LM) TV program. We discarded most data in this TV program after 2-pass data cleaning. It happened because this particular TV show was quite challenging for our ASR model.

3. Hybrid speech recognition

3.1. Acoustic Model

In all our experiments, the acoustic model was based on a hybrid Deep Neural Network – Hidden Markov Model architecture trained in Kaldi [12]. The NN part of the model contains 6 convolutional layers followed by 19 TDNN layers with semi-orthogonal factorization [1] (CNN-TDNNf). The input consists of 40-dim MFCCs concatenated with speaker dependent 100-dim i-vectors. Whole model is trained using LF-MMI objective function with bi-phone acoustic units as the targets.

In order to make our NN model training more robust, we introduced feature dropout layer into the architecture. This prevents the model from overfitting on training data. In fact, it turned the overfitting problem into an underfitting problem. Thus, it leads to a slower convergence during training. This is solved by increasing the number of epochs from 6 to 8 to balance the underfitting in our system. This technique is also known as Spectral Augmentation. It was first suggested for multi-stream hybrid NN models in [13] and fully examined in [14].

3.2. Language Model

We trained three different 3-gram language models: Alb, Wiki and Giga. The names suggest which text corpus was used during training. Albayzin LM was trained on dev1 and train sets from RTVE2018. This text mixture contains 80 thousand unique words in 0.5 million sentences. This small training text is not optimal to train N-gram LM, which is able to generalize well. So we also included larger text corpora: Wikipedia and Spanish Gigaword. These databases were further processed to get rid of unrelated text such as advertisements, emojis and urls. This resulted into more than 2.5 million fine sentences in Wikipedia and 20 million sentences in Spanish Gigaword. We experimented with 4 combinations of interpolation: Alb, Alb+Wiki, Alb+Giga, Alb+Wiki+Giga.

Our vocabulary consists of words from the RTVE2018 database and from the Santiago lexicon[1]. The pronunciation of Spanish words was extracted using a public TTS model called E-speak [15]. The vocabulary was then extended by auxiliary labels for noise, music and overlapped speech. The final lexicon contains around 110 thousand words.

3.3. Voice Activity Detection

Voice activity detection (VAD) was applied on evaluation data in order to segment the audio into smaller chunks. The VAD approach used in this system is based on a feed-forward neural network with two outputs. It expects 15-dimensional filterbank features with additional 3 Kaldi pitch features [16] as the input. Features are normalized with cepstral mean normalization. More details can be found in [17].

4. End-to-end speech recognition

4.1. Acoustic Model

The end-to-end acoustic model is based on a convolutional architecture proposed by Collobert et al. [2] that uses gated linear units (GLUs). Using GLUs in convolutional approaches helps avoiding vanishing gradients, by providing them linear paths while keeping high performances. Concretely, we have used the model from wav2letter Wall Street Journal (WSJ) recipe. This model has approximately 17M parameters with dropout applied after each of its 17 layers. The WSJ dataset contains around 80 hours of audio recordings, which is smaller than the magnitude of our data (∼600 hours). The LibriSpeech recipe (∼1000 hours) provides a deeper ConvNet GLU based architecture, however we decided to use the WSJ one in order to reduce computational time and improve hyper-parameter fine-tuning of the network.

All data samples are resampled at 16kHz, and the system is trained with wav2letter++ framework. Mel-frequency spectral coefficients (MFSCs) are extracted from raw audio, using 80 filterbanks, and the system is trained using the Auto Segmentation criterion (ASG) [2] with batch size set to 4. The learning

[1] https://www.openslr.org/34/

rate starts at 5.6 and is decreased down to 0.4 after 30 epochs, where training is finished since no significant WER gains are achieved. From epochs 22 to 28 the system is trained also with the same train set, but adding the RTVE2018 train and dev1 samples with the background music cleaned by Demucs module [18]. The last two epochs, from epoch 28 to epoch 30, are done incorporating further samples with background noise removed by Demucs and denoised by a neural denoiser [19]. This way, data augmentation with samples without background music and noise is done, to aid the network at training with samples with difficult acoustic conditions. Besides, the network is more likely to generalize audio artifacts caused by the denoiser and music separator networks, which is useful when using these to clean test audio.

4.2. Language Model

Regarding the lexicon, we extract it from the train and validation transcripts, plus Sala lexicon [20]. The resulting lexicon is a grapheme-based one with 271k words. We use the standard Spanish alphabet as tokens, plus the "ç" letter from some Catalan words found in the training data, plus the vowels with diacritical marks, making a total of 37 tokens.

The LM is a 5-gram model trained with KenLM [21] using only transcripts from the training sets: RTVE2018 train and dev1, plus Common Voice, Fisher and Call Center. The resulting LM is described in this paper as *Alb+Others*.

Fine-tuning of decoder hyperparameters is done via grid-search with the RTVE2018 dev2 set. The best results are achieved with an LM weight of 2.25, a word score of 2.25 and a silence score of -0.35. This same configuration is then applied for evaluation datasets from RTVE2018 and RTVE2020.

5. Experiments

5.1. Data cleaning

Data cleaning by means of 2-pass transcript retrieval improves the performance of our models the most. Table 2 shows the effect of each pass. The 2nd pass helped to improve the accuracy by almost 2 % in terms of WER. We also ran a 3rd pass, but that did not help anymore. We simply did not retrieve more cleaned data from the original transcripts, just 3 hours more.

Table 2: *Effect of 2-pass transcript cleaning evaluated on RTVE2018 test set.*

AM	LM	Training data	WER [%] Test
CNN-TDNNf	Alb	1-pass	17.2
		2-pass	15.5
		3-pass	15.5

5.2. Speech Enhancement

It is very common to find background music on TV programs, which can confuse our recognizer if it has a notorious presence. This brought us the idea of processing the audio through a Music Source Separator called Demucs [18]. It separates the original audio into voice, bass, drums and others. By keeping only the voice component, we managed to significantly eliminate the background music, while maintaining relatively good quality in the original voice.

We enhanced both validation sets in order to assess possible WER reductions. As seen in Table 4, this approach yielded a small increase in WER. We also tried applying a specialized denoiser [19] after background music removal, but the WER for dev2 increased in an absolute 1.6%, compared to original system without enhancement. None of these two approaches (Demucs and Demucs+Denoiser) provided WER improvements at first, so we did not apply them for the end-to-end model used in the fusion. Nevertheless, the end-to-end, end-to-end + Demucs and end-to-end + Demucs + Denoiser models were submitted as separate systems by the UPF-TID team (see Table 5 for details).

Our hypothesis is that not all the samples contain background music. Speech enhancement for samples that are clean already is detrimental because it causes slight degradations in the signal. Hence, we have evaluated the effects of applying music source separation to samples under certain SNR ranges, measured with the WADA-SNR algorithm [22]. The application of music separation on RTVE dataset is optimal for SNR ranges between -5 and 5 or 8 as it is shown in Table 3. Looking at Figure 2, best improvements are found at TV shows with higher WER (thus harder/noisier speech), e.g., AV, where most of the time speakers are in a car, or LM and DH, where music and speech often overlap. Other shows have slighter benefits, since these already contain good quality audio. The exception is AFI show, which is reported to have poor quality audio, so further audio degradation from Demucs might cause worse performance.

Figure 2: *Variation of the mean WER per TV show between using Demucs-cleaned or original samples on RTVE's 2018 test set. Negative values represent Demucs improvements. Note that only samples with SNR between -5 and 8 are enhanced.*

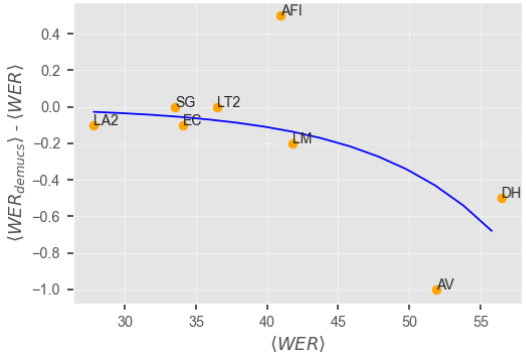


Table 3: *WER impact of cleaning speech signals between certain SNR ranges, using a music source separator. End-to-end ConvNet GLU model is used without LM, and percentage of cleaned samples are reported.*

SNR	Cleaned Samples [%] 2018	Cleaned Samples [%] 2020	Test WER [%] 2018	Test WER [%] 2020
$(-\infty, \infty)$	100	100	37.50	53.53
$(-\infty, 10)$	25.97	34.22	-0.05	-0.87
$(-5, 10)$	25.84	31.33	-0.05	-0.88
$(-5, 5)$	5.14	11.88	-0.07	**-1.03**
$(-5, 8)$	14.95	22.11	**-0.08**	-0.97

5.3. Spectral augmentation

Table 4 shows compared models with and without spectral augmentation. The technique helps quite significantly. All models with feature dropout layer outperformed their counterparts with a quite constant 0.4% absolute WER improvement on RTVE2018 test set and around 0.6% on RTVE2018 dev2 set.

5.4. Model fusion

We also fuse the output of our best systems to further improve the performance. Overall results of our systems considered for the fusion are depicted in Table 4. Since the models with spectral augmentation performed significantly better, we decided to fuse only these systems. We analyzed two different approaches: a pure hybrid model fusion (Fusion A) and hybrid and end-to-end model fusion (Fusion B).

Considering that the end-to-end model does not provide word-level timestamps, we had to force-align the transcripts with the hybrid ASR system in order to obtain CTM output. The original word-level fusion was done using ROVER toolkit [23]. Fusion B with end-to-end models performed slightly better than its counterpart Fusion A, despite the fact that the end-to-end models achieved worse results. This somehow proves the idea that the fusion can benefit from different modeling approaches.

6. Final systems

Table 5 depicts the results on RTVE2020 test set. For the end-to-end ConvNet GLU model, the performance drops around a 15% WER when compared with previous results on development sets. Since the TV shows in such sets are also present in training dataset, our hypothesis is that the model slightly overfits to them. Therefore, when facing different acoustic conditions, voices, background noises and music types included in the RTVE2020 test set, the WER increases noticeably. Enhancing the test samples with Demucs or with Demucs+Denoiser yields a worse WER score, probably due to an inherent degradation of the signal. A deeper analysis about more efficient ways to apply such enhancements has been done in Section 5.2.

Also, note that the submitted systems had a leak of dev2 stm transcripts in the LM, causing an hyperparameter overfitting during LM tuning. This caused a WER drop in all end-to-end systems, yielding WERs of 41.4%, 42.3% and 58.6%. Table 5 also displays the results of these systems with the leakage and LM tuning corrected in post-evaluation analysis.

7. Conclusions

In this paper we described two different ASR model architectures and their fusion. We focused on improving the original subtitled data in order to train our models on high quality target labels. We also improved the N-gram language model by incorporating publicly available text data from Wikipedia and Spanish Gigaword corpus from LDC. We have also successfully incorporated the spectral augmentation into our AM architecture. Our best system achieved 13.3 % and 23.24 % WER on RTVE2018 and RTVE2020 test sets respectively.

The performance of our hybrid system can be further improved by using lattice-fusion with Minimum Bayes Risk decoding [24]. Another space for improvement is offered by adding an RNN-LM lattice rescoring. Our end-to-end model shows relatively competitive performance on the RTVE2018 test set in comparison with its hybrid counterpart. However, its performance on the RTVE2020 evaluation set exposes that the model was not able to generalize very well since this database turns out to contain slightly different acoustic conditions. Despite of this fact, the model still managed to improve the results in the final fusion with hybrid systems. An exploration on background music removal shows that it yields the best results for lower SNR ranges, thus having a different impact depending on the acoustic conditions of each TV show.

Table 4: *Overall results on RTVE2018 dataset with various language models and fusions.*

	AM	LM	WER [%] Dev2	WER [%] Test
1	CNN-TDNNf	Alb	14.1	15.5
2		Alb + Wiki	13.6	14.9
3		Alb + Giga	13.6	15.1
4		Alb + Wiki + Giga	13.5	15.0
5	+ SpecAug	Alb	13.4	15.0
6		Alb+Wiki	12.9	14.5
7		Alb+Giga	13.0	14.7
8		Alb+Wiki+Giga	12.9	14.6
9	ConvNet GLU	None	36.1	37.5
10		Alb + Others	20.8	20.7
11	+ Demucs	None	36.4	37.5
12		Alb + Others	21.1	20.8
13	Fusion A	*(row 5-8)*	12.9	13.7
14	Fusion B	*(row 5-8 and 10)*	**12.8**	**13.3**

Table 5: *Official and post-evaluation final results on RTVE2020 eval set for the submitted systems.*

Model	WER [%] Official	WER [%] Post-eval
CNN-TDNNf	-	24.3
+ SpecAug	-	23.5
ConvNet GLU	41.4[2]	36.2
+ Demucs	42.3[3]	37.9
+ Demucs + Denoiser	58.6[4]	40.0
Fusion A	23.33	23.38
Fusion B	**23.33**	**23.24**

8. Acknowledgements

The work was partly supported by Czech National Science Foundation (GACR) project NEUREM3 No. 19-26934X, European Union's Horizon 2020 project No. 833435 - INGENIOUS and Horizon 2020 project No. 864702 - ATCO2, and by Czech Ministry of Education, Youth and Sports project no. LTAIN19087 "Multi-linguality in speech technologies". Mireia Farrús has been funded by the Agencia Estatal de Investigación (AEI), Ministerio de Ciencia, Innovación y Universidades and the Fondo Social Europeo (FSE) under grant RYC-2015-17239 (AEI/FSE, UE).

[2]Primary system of UPF-TID team.
[3]First contrastive system of UPF-TID team.
[4]Second contrastive system of UPF-TID team.

9. References

[1] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Proceedings of Interspeech*, 09 2018, pp. 3743–3747.

[2] R. Collobert, C. Puhrsch, and G. Synnaeve, "Wav2letter: an end-to-end convnet-based speech recognition system," *CoRR*, vol. abs/1609.03193, 2016. [Online]. Available: http://arxiv.org/abs/1609.03193

[3] E. Lleida, A. Ortega, A. Miguel, V. Bazán, C. Pérez, M. Zotano, and A. De Prada, "RTVE2018 Database Description," 2018. [Online]. Available: http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf

[4] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. de Prada, "Albayzin 2018 evaluation: the IberSPEECH-RTVE challenge on speech technologies for Spanish broadcast media," *Applied Sciences*, vol. 9, no. 24, p. 5412, 2019.

[5] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. De Prada, "RTVE2020 Database Description," 2020. [Online]. Available: http://catedrartve.unizar.es/reto2020/RTVE2020DB.pdf

[6] D. Graff, S. Huang, I. Cartagena, K. Walker, and C. Cieri, "Fisher Spanish Speech," *LDC2010S01. DVD. Philadelphia: Linguistic Data Consortium*, 2010.

[7] A. Canavan and G. Zipperlen, "CALLHOME Spanish Speech," *LDC96S35. Web Download. Philadelphia: Linguistic Data Consortium*, 1996.

[8] Ângelo Mendonça, D. Jaquette, D. Graff, and D. DiPersio, "Spanish Gigaword Third Edition," *LDC2011T12. Web Download. Philadelphia: Linguistic Data Consortium*, 2011.

[9] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," 2019.

[10] V. Manohar, D. Povey, and S. Khudanpur, "JHU Kaldi system for Arabic MGB-3 ASR challenge using diarization, audio-transcript alignment and transfer learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, vol. 2018-. IEEE, 2017, pp. 346–352.

[11] M. Kocour, "Automatic Speech Recognition System Continually Improving Based on Subtitled Speech Data," Diploma thesis, Brno University of Technology, Faculty of Information Technology, Brno, 2019, technical supervisor Dr. Ing. Jordi Luque Serrano. supervisor Doc. Dr. Ing. Jan Černocky.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[13] S. H. R. Mallidi and H. Hermansky, "A Framework for Practical Multistream ASR," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, N. Morgan, Ed. ISCA, 2016, pp. 3474–3478.

[14] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*. ISCA, 2019, pp. 2613–2617.

[15] J. Duddington and R. Dunn, "eSpeak text to speech," 2012. [Online]. Available: http://espeak.sourceforge.net

[16] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. Florence, Italy: IEEE, May 2014.

[17] O. Plchot, P. Matějka, O. Novotný, S. Cumani, A. D. Lozano, J. Slavíček, M. S. Diez, F. Grézl, O. Glembek, M. V. Kamsali, A. Silnova, L. Burget, L. Ondel, S. Kesiraju, and A. J. Rohdin, "Analysis of BUT-PT Submission for NIST LRE 2017," in *Proceedings of Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 47–53.

[18] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.

[19] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," 2020.

[20] A. Moreno, O. Gedge, H. Heuvel, H. Höge, S. Horbach, P. Martin, E. Pinto, A. Rincón, F. Senia, and R. Sukkar, "SpeechDat across all America: SALA II," 2002.

[21] K. Heafield, "KenLM: Faster and Smaller Language Model Queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, ser. WMT '11. USA: Association for Computational Linguistics, 2011, p. 187–197.

[22] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[23] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, 1997, pp. 347–354.

[24] P. Swietojanski, A. Ghoshal, and S. Renals, "Revisiting hybrid and GMM-HMM system combination techniques," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, 2013.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

MLLP-VRAIN Spanish ASR Systems for the Albayzin-RTVE 2020 Speech-To-Text Challenge

Javier Jorge, Adrià Giménez, Pau Baquero-Arnal, Javier Iranzo-Sánchez, Alejandro Pérez, Gonçal V. Garcés Díaz-Munío, Joan Albert Silvestre-Cerdà, Jorge Civera, Albert Sanchis and Alfons Juan

Machine Learning and Language Processing (MLLP) research group
Valencian Research Institute for Artificial Intelligence (VRAIN)
Universitat Politècnica de València
Camí de Vera s/n, 46022, València, Spain

{jajorca,adgipas,pabaar,jairsan,alpegon2,gogardia,
juasilce,jorcisai,josanna2,ajuanci}@vrain.upv.es

Abstract

This paper describes the automatic speech recognition (ASR) systems built by the MLLP-VRAIN research group of Universitat Politècnica de València for the Albayzin-RTVE 2020 Speech-to-Text Challenge.

The primary system (*p-streaming_1500ms_nlt*) was a hybrid BLSTM-HMM ASR system using streaming one-pass decoding with a context window of 1.5 seconds and a linear combination of an n-gram, a LSTM, and a Transformer language model (LM). The acoustic model was trained on nearly 4,000 hours of speech data from different sources, using the MLLP's transLectures-UPV toolkit (TLK) and TensorFlow; whilst LMs were trained using SRILM (n-gram), CUED-RNNLM (LSTM), and Fairseq (Transformer), with up to 102G tokens. This system achieved 11.6% and 16.0% WER on the *test-2018* and *test-2020* sets, respectively. As it is streaming-enabled, it could be put into production environments for automatic captioning of live media streams, with a theoretical delay of 1.5 seconds.

Along with the primary system, we also submitted three contrastive systems. From these, we highlight the system *c2-streaming_600ms_t* that, following the same configuration of the primary one, but using a smaller context window of 0.6 seconds and a Transformer LM, scored 12.3% and 16.9% WER points respectively on the same test sets, with a measured empirical latency of 0.81±0.09 seconds (mean±stdev). This is, we obtained state-of-the-art latencies for high-quality automatic live captioning with a small WER degradation of 6% relative.

Index Terms: natural language processing, automatic speech recognition, streaming.

1. Introduction

This paper describes the participation of the *Machine Learning and Language Processing* (MLLP) research group from the *Valencian Research Institute for Artificial Intelligence* (VRAIN), hosted at the *Universitat Politècnica de València* (UPV), in the Albayzin-RTVE 2020 Speech-to-Text (S2T) Challenge.

Live audio and video streams such as TV broadcasts, conferences, lectures, as well as general-public video streaming services (e.g. Youtube) over the Internet have increased dramatically in recent years because of the advances in networking with high speed connections and proper bandwidth. Also, due to the COVID-19 pandemic, video meeting/conferencing platforms have experienced an exponential growth of usage, as public and private companies have leveraged teleworking for their employees to comply with the social distancing measures recommended by health authorities.

Automatic transcription and translation of such audio streams is a key feature in a globalized and interconnected world, in order to reach wider audiences or to ensure proper understanding between native and non-native speakers, depending on the use-case. Also, public governments are enforcing TV broadcasters by law to provide accessibility options to people with hearing disabilities, with a yearly increasing amount of contents to be captioned at a minimum [1, 2].

Some TV broadcasters and other live streaming services have assumed manual transcription from scratch of live audio or video streams, as an initial solution to comply with the current legislation, and/or to satisfy user expectations. However, it is a really hard task for professional linguists that, under very stressful conditions, are very prone to generate captioning errors. Besides, it is difficult to scale up such a service, as in these organizations, the amount of human resources devoted to this particular task is typically scarce.

Due to these reasons, the need and demand for high-quality real-time streaming Automatic Speech Recognition (ASR) has increased drastically in the last years. Automatic live audio stream subtitling enables professional linguists to correct live transcripts provided by these ASR systems, if they are not publishable as they come. This would dramatically expedite their productivity and significantly reduce the probability of producing transcription errors. However, the application of state-of-the-art ASR technology to video streaming is a highly complex and challenging task due to real-time and low-latency recognition constraints.

The MLLP-VRAIN, being aware of these demands from the society, have focused its research efforts in the past two years on streaming ASR. This work aims to disseminate our latest developments in this area, showing how our hybrid ASR technology can be successfully applied under streaming conditions, by providing high-quality transcriptions and state-of-the-art system latencies on real-life tasks such as the RTVE (*Radio Televisión Española*) database. Therefore, our participation in the Albayzin-RTVE 2020 S2T Challenge consisted on the submission of a primary, performance-focused streaming ASR system, plus three contrastive systems: two latency-focused streaming ASR systems, and one conventional off-line ASR system.

10.21437/IberSPEECH.2021-25

Table 1: *Transcribed Spanish speech resources for AM training.*

Resource	Duration (h)
Internal: entertainment	2932
Internal: educational	406
Internal: user-generated content	202
Internal: parliamentary data	158
Voxforge [8]	21
RTVE2018: *train*	187
RTVE2018: *dev1-train*	18
TOTAL	3924

The rest of the paper is structured as follows. First, Section 2 briefly describes the Albayzin-RTVE 2020 S2T Challenge and the RTVE databases provided by the organizers. Next, Section 3 provides a detailed description of our participant ASR systems. Finally, Section 4 gives a summary of the work plus some concluding remarks.

2. Challenge description and databases

The Albayzin-RTVE 2020 Speech-To-Text Challenge consists of automatically transcribing different types of TV shows from the RTVE Spanish public TV station, and the assessment of ASR system performance in terms of Word Error Rate (WER) by comparing those automatic transcriptions with correct reference transcriptions [3].

The MLLP-VRAIN participated in the 2018 edition of the challenge [4] in a joint collaboration with the *Human Language Technology and Pattern Recognition* (HLTPR) research group from the *RWTH Aachen University*. The evaluation was carried out on the RTVE2018 database [5], that includes 575 hours of audio from 15 different TV shows broadcasted between 2015 and 2018. This database is allocated into four sets: *train, dev1, dev2* and *test* (*test-2018*). Our systems won in both the open-condition and closed-condition tracks [6], scoring 16.5% and 22.0% WER points respectively in the *test-2018* set.

For the 2020 edition of the challenge, the participation has been limited to a single open-condition track, and system evaluations have been carried out over the *test* (*test-2020*) set from the RTVE2020 database, which includes 78.4 hours from 15 different TV shows broadcasted between 2018 and 2019 [7].

3. MLLP-VRAIN Systems

In this section we describe the hybrid ASR systems developed by the MLLP-VRAIN that participated in the Albayzin-RTVE 2020 S2T Challenge.

3.1. Acoustic Modelling

Our acoustic models (AM) were trained using 205 filtered speech hours from the *train* set (187h) and our internal *dev1-train* set (18h), as in [4], plus about 3.7K hours of other resources crawled from the Internet. Table 1 summarises all training datasets along with their total duration (in hours). From this data, first, we extracted 16-dimensional MFCC features plus first and second derivatives (48-dimensional feature vectors) every 10ms to train a context-dependent feed-forward DNN-HMM with three left-to-right tied states using the transLectures-UPV toolkit (TLK) [9]. The state-tying schema followed a phonetic decision tree approach [10] that produced 10K tied states. Then, feed-forward models were used to bootstrap a BLSTM-HMM AM, trained with 85-dimensional filterbank features, following the procedure described in [11]. The BLSTM network was trained using both TLK and TensorFlow [12], and had 8 bidirectional hidden layers with 512 LSTM cells per layer and direction. As in [11], we performed chunking during training by considering a context to perform back-propagation through time to a window size of 50 frames. Additionally, SpecAugmentation was applied by means of time and frequency distortions [13].

3.2. Language Modelling

Regarding language modelling, we trained count-based (n-gram) and neural-based (LSTM, Transformer) Language Models (LMs) to perform one-pass decoding with different linear combinations of them [14], using the text data sources and corpora described in Table 2.

On the one hand, we trained 4-gram LMs using SRILM [15] with all text resources plus the Google-counts v2 corpus [16], accounting for 102G running words. The vocabulary size was limited to 254K words, with an OOV ratio of 0.6% computed over our internal development set.

On the other hand, regarding neural LMs, we considered the LSTM and Transformer architectures. In both cases, LMs were trained using a 1-gigaword subset randomly extracted from all available text resources, except Google-counts. Their vocabulary was defined as the intersection between the n-gram vocabulary (254K words) and that derived from the aforementioned training subset. We did this to avoid having zero probabilities for words that are present in the system vocabulary but not in the training subset. This is taken into account when computing perplexities by renormalizing the unknown-word score accordingly.

Specific training details for each neural LM architecture are as follows. Firstly, LSTM LMs were trained using the CUED-RNNLM toolkit [17]. Noise Contrastive Estimation (NCE) criterion [18] was used to speed up model training, and the normalization constant learned from training was used during decoding [19]. Based on the lowest perplexity observed on our internal development set, we selected as final model that with a 256-unit embedding layer and two hidden LSTM layers of 2048 units. Secondly, Transformer LMs (TLMs) were trained using a customized version of the FairSeq toolkit [20], selecting the following configuration that minimized perplexity in our internal development set: 24-layer network with 768 units per layer, 4096-unit FFN, 12 attention heads, and an embedding of 768 dimensions. These models were trained until convergence with batches limited to 512 tokens, 512 sentences, and 512 words per sentence. Parameters were updated every 32 batches. During inference, Variance Regularization (VR) was applied to speed up the computation of the TLM score [21].

3.3. Decoding strategy

Our hybrid ASR systems follow a real-time one-pass decoding by means of a History Conditioned Search (HCS) strategy, as described in [14]. This approach allows us to benefit from the direct usage of additional LMs during decoding while satisfying real-time constraints. This decoding strategy introduces two additional and relevant parameters to control the trade-off between Real Time Factor (RTF) and WER: LM history recombination (LMHR), and LM histogram prunning (LMHP). The static look-ahead table, needed by the decoder to use precomputed look-ahead LM scores, was generated from a prunned

Table 2: *Statistics of Spanish text resources for LM training. S=Sentences, RW=Running words, V=Vocabulary. Units are in thousands (K).*

Corpus	S(K)	RW(K)	V(K)
Opensubtitles [22]	212635	1146861	1576
UFAL [23]	92873	910728	2179
Wikipedia [24]	32686	586068	3373
UN [25]	11196	343594	381
News Crawl [26]	7532	198545	648
Internal: entertainment	4799	59235	307
eldiario.es [27]	1665	47542	247
El Periódico [28]	2677	46637	291
Common Crawl [29]	1719	41792	486
Internal: parliamentary data	1361	35170	126
News Commentary [26]	207	5448	83
Internal: educational	87	1526	35
TOTAL	369434	3423146	5785
Google-counts v2 [16]	-	97447282	3693

Table 3: *Basic statistics of development and tests sets of RTVE databases, including our internal dev1-dev set: total duration (in hours), number of files, average duration of samples in seconds plus-minus standard deviation ($d_\mu \pm \sigma$), and running words (RW) in thousands (K).*

Set	Duration(h)	Files	d_μ	$\pm$ σ	RW(K)
dev1-dev	11.9	10	4267	$\pm$ 1549	120
dev2	15.2	12	4564	$\pm$ 1557	149
test-2018	39.3	59	2395	$\pm$ 1673	377
test-2020	78.4	87	2314	$\pm$ 1576	519

version of the n-gram LM.

For streaming ASR, as the full sequence (context) is not available during decoding, BLSTM AMs are queried with a sliding, overlapping context window of limited size over the input sequence, averaging outputs of all windows for each frame to obtain the corresponding acoustic score [30]. The size of the context window (in frames or seconds) is set in decoding, and defines the theoretical latency of the system. This limitation of the context prevents us to perform a Full Sequence Normalization (FSN), that is typically applied under the off-line setting. Instead, we applied the Weighted Moving Average (WMA) technique, that uses the content of the current context window to update normalization statistics on-the-fly, weighted by previous context from past windows with an α parameter [31]. Finally, as Transformer LMs have the inherent capacity of attending to potentially infinite word sequences, history is limited to a given maximum number of words, in order to meet the strict computational time constraints imposed by the streaming scenario [21]. By applying all these modifications, our decoder acquires the capacity to deliver live transcriptions for incoming audio streams of potentially infinite length, with latencies lower-bounded by the context window size.

3.4. Experiments and results

To carry out our experiments, we used the development and test sets from the RTVE2018 database. More precisely, we devoted our internal *dev1-dev* set [4] for development purposes, whilst *dev2* and *test-2018* were dedicated to test ASR performance. Finally, *test-2020* was the blind test used by the organisation to rank the participant systems. Table 3 provides basic statistics of these sets.

Table 4: *Perplexity (PPL) and interpolation weights, computed over the dev1-dev set, of all possible linear combinations of n-gram (ng), LSTM (ls) and Transformer (tf) LMs.*

LM comb.	PPL	Weights(%)
ng	179.5	-
ls	98.4	-
tf	63.3	-
ng + ls	93.2	15 + 85
ng + tf	61.6	6 + 94
ls + tf	60.7	13 + 87
ng + ls + tf	59.5	5 + 10 + 85

First, we studied the perplexity (PPL) on the *dev1-dev* set of all possible linear combinations for the three types of LMs considered in this work. Table 4 shows the PPLs of these interpolations, along with the optimum LM weights that minimized PPL in the *dev1-dev* set. The Transformer LM provides significant lower perplexities in all cases, and accordingly, takes very high weight values when combined with other LMs. Indeed, the TLM in isolation already delivers a strong perplexity baseline value of 63.3, while the maximum PPL improvement is of just 6% relative when all three LMs are combined.

Second, we tuned decoding parameters to provide a good WER-RTF tradeoff on *dev1-dev*, with the hard constraint of RTF$<$1 to ensure a real-time processing of the input. From these hiperparameters, we highlight, due to their relevance, LMHR=12, LMHP=20, and TLM history limited to 40 words.

At this point, we defined our participant off-line hybrid ASR system identified as *c3-offline* (contrastive system no. 3), consisting of a fast pre-recognition + Voice Activity Detection (VAD) step to detect speech/no-speech segments as in [4], followed by a real-time one-pass decoding with our BLSTM-HMM AM, using a FSN normalization scheme and a linear combination of the three types of LMs: n-gram, LSTM and Transformer. This system scored 12.3 and 17.1 WER points on test-2018 and test-2020, respectively.

Next, as our focus was to develop the best-performing streaming-capable hybrid ASR system for this competition, we explored streaming-related decoding parameters to optimize WER on *dev1-dev*, using the BLSTM-HMM AM and a linear combination of all three LMs. This resulted on using a context window size of 1.5 seconds and α=0.95 for the WMA normalization technique. This configuration defined our primary system, identified as *p-streaming_1500ms_nlt*, that showed WER rates of 11.6 and 16.0 in *test-2018* and *test-2020*, respectively. It is important to note that this system does not integrate any VAD module. This task is implicitly carried out by the decoder via the non-speech model of the BLSTM-HMM AM.

A small change on the configuration of the primary system, consisting on the removal of the LSTM LM from the linear interpolation, defined the contrastive system no. 1, identified as *c1-streaming_1500ms_nt*. The motivation behind this change is that the computation of LSTM LM scores is quite expensive in computational terms, and its contribution to PPL is negligible with respect to the n-gram LM + TLM combination (3% relative improvement). Hence, for the sake of system latency stability, we obtained nearly no degradation in terms of WER: 11.6 and 16.1 points in *test-2018* and *test-2020*, respectively.

Both streaming ASR systems, *p-streaming_1500ms_nlt* and *c1-streaming_1500ms_nt*, share the same theoretical latency of 1.5 seconds, as it is determined by the context window size. As

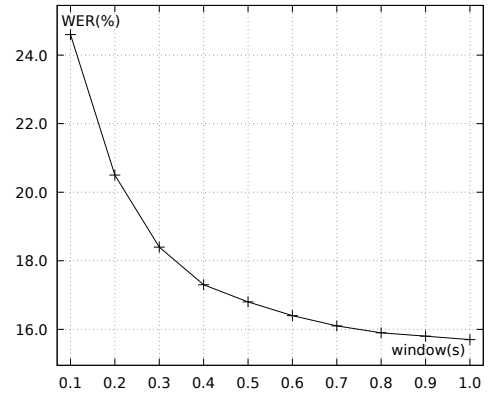


Figure 1: *WER as a function of context window size (in seconds) for the streaming setup, computed over the dev1-dev set.*

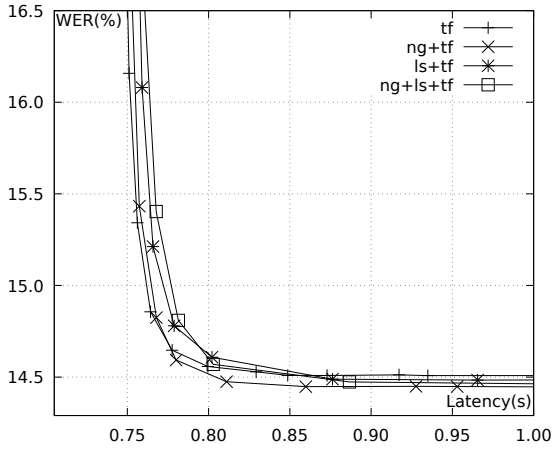


Figure 2: *WER versus mean empirical latency (in seconds) on dev1-dev, measured with different prunning parameters, and considering only interpolation schemes that include TLM.*

Table 5: *WER of the participant systems, including our open-condition system submitted to the 2018 challenge, computed over the dev2, test-2018 and test-2020 sets.*

System	*dev2*	*test-2018*	*test-2020*
p-streaming_1500ms_nlt	11.2	11.6	16.0
c1-streaming_1500ms_nt	-	11.6	16.1
c2-streaming_600ms_t	12.0	12.3	16.9
c3-offline	-	12.0	17.1
2018 open-cond. winner [4]	15.6	16.5	-

stated in Section 3.3, this parameter can be adjusted in decoding time. This allows us to configure the decoder for lower latency responses or better transcription quality. Hence, our last commitment for this challenge was to find a proper system configuration that could provide state-of-the-art, stable latencies with minimal WER degradation. Figure 1 illustrates the evolution of WER on *dev1-dev* as a function of the context window size, limited to one second at maximum. As we focused on gauging AM performance, we used the n-gram LM in isolation for efficiency reasons. At the light of the results, we chose a window size of 0.6 seconds, as it brings a good balance between transcription quality and theoretical latency.

The last step to set up our latency-focused streaming system was to measure WER and empirical latencies as a function of different prunning parameters and LM combinations. In our experiments, latency is measured as the time elapsed between the instant at which an acoustic frame is generated, and the instant at it is fully processed by the decoder. We provide latency figures at the dataset level, computed as the average of the latencies observed at the frame level on the whole dataset. Figure 2 shows WER vs mean empirical latency figures, computed over *dev1-dev*, with different prunning parameter values, and comparing the LM combinations that include the Transformer LM. These measurements were run on an Intel i7-3820 CPU @ 3.60GHz, with 64GB of RAM and a RTX 2080 Ti GPU card. On the one hand, we can see how combinations involving LSTM LMs are systematically shifted rightwards w.r.t. other combinations. This means that the LSTM LM has a clear negative impact on system latency, with little to no effect on system quality. This evidence corroborates our decision of removing the LSTM LM to define our contrastive system *c1-streaming_1500ms_nt*. On the other hand, TLM alone generally provides a good baseline that is slightly improved in terms of WER if we include the other LMs. However, this comes with the cost of increasing latency. Hence, we selected the Transformer LM in isolation for our final latency-focused streaming system. This system was our contrastive system no. 2, identified as *c2-streaming_600ms_t*. Its empirical latency on *dev1-dev* was 0.81 ± 0.09 seconds (mean$\pm$stdev), and its performance was 12.3 and 16.9 WER points in *test-2018* and *test-2020*, respectively. This is, with just a very small relative WER degradation of 6% w.r.t. the primary system, we got state-of-the-art (mean=0.81s) and very stable (stdev=0.09s) empirical latencies. This system has a baseline consumption (when idle) of 9GB RAM and 3.5GB GPU memory (on a single GPU card), adding 256MB RAM and one CPU thread per each decoding (audio stream). For instance, the decoding of four simultaneous audio streams in a single machine would use four CPU threads, 10GB RAM and 3.5GB GPU memory.

Table 5 summarises the results obtained with all the four participant ASR systems in the *dev2*, *test-2018* and *test-2020* sets, and adds the results obtained with our 2018 open-condition system for comparison. On the one hand, surprisingly, the offline system is surpassed by the three streaming ones in *test-2020*, by up to 1.1 absolute WER points (6% relative). We believe that this is caused, first, by an improvable VAD module, based on Gaussian Mixture HMMs, that, in our experience, suffers from false negatives (speech segments labelled as non-speech). As the non-speech model was trained with music and noise audio segments, and given the inherent limitations of GMMs, it is likely to misclassify speech passages with loud background music and noise (often present in TV programmes) as non-speech. Second, the FSN technique might not be appropriate for some types of TV shows, as local acoustic condition changes become diluted in the full-sequence normalization, and acoustic scores computed for those frames may present some perturbations that can degrade system performance at that point. On the other hand, it is remarkable that our primary 2020 system significantly outperforms the 2018 winning system by 28% relative WER points on both *dev2* and *test-2018* (25% in the case of our latency-focused system *c2-streaming_600ms_t*), while adding the novel streaming capability at the same time.

All these streaming ASR systems can be easily put into production environments using our custom gRPC-based server-

client infrastructure[1]. Indeed, ASR systems comparable to *c2-streaming_600ms_t* and *c1-streaming_1500ms_nt* are already in production at our Transcription and Translation Plarform (TTP)[2] for streaming and off-line processing, respectively. Both can be freely tested using our public APIs, accessible via TTP.

4. Conclusions

In this paper we have described our four ASR systems that participated in the Albayzin-RTVE 2020 Speech-to-Text Challenge. The primary one, a streaming-enabled performance-focused hybrid ASR system (*p-streaming_1500ms_nlt*) provided a good score of 16.0 WER points in the *test-2020* set, and a remarkable 28% relative WER improvement over the 2018 winning ASR system on *test-2018*, with a theoretical latency of 1.5 seconds. Nearly the same performance was delivered by our first contrastive system (*c1-streaming_1500ms_nt*): 16.1 WER points on *test-2020*, at a significant lower computational cost. In pursuit of low, state-of-the-art system latencies, our second contrastive system (*c2-streaming_600ms_t*) provided a groundbreaking WER-latency balance, with a solid performance of 16.9 WER points on *test-2020* at an empirical latency of 0.81 ± 0.09 seconds (mean$\pm$stdev). Finally, our contrastive off-line ASR system with VAD (*c3-offline*) provides the highest, yet still competitive, WER mark of 17.1 points, attributable to an improvable VAD module and to the limitations of FSN when dealing with local acoustic condition changes.

With a configurable system latency in decoding time, our ASR technology offers the flexibility to produce fast system responses for streaming applications, or to generate maximum quality transcriptions whenever hard time constraints do not apply. Also, results demonstrate that our streaming ASR technology is mature enough to be systematically put into production environments for high-quality automatic live captioning in TV stations, distance learning, conferencing platforms, or general-purpose video/audio streaming services, among others.

5. Acknowledgements

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 761758 (X5Gon); the Government of Spain's research project Multisub (ref. RTI2018-094879-B-I00, MCIU/AEI/FEDER,EU) and FPU scholarships FPU14/03981 and FPU18/04135; and the Generalitat Valenciana's research project Classroom Activity Recognition (ref. PROMETEO/2019/111) and predoctoral research scholarship ACIF/2017/055.

6. References

[1] "RD 1494/2007, del 12 de noviembre," 2007. [Online]. Available: https://www.boe.es/buscar/act.php?id=BOE-A-2007-19968

[2] "Llei 1/2006, de 19 d'abril, GVA," 2006. [Online]. Available: https://www.dogv.gva.es/va/eli/es-vc/l/2006/04/19/1/dof/vci-spa/pdf

[3] E. Lleida *et al.*, "IberSPEECH-RTVE 2020 speech to text transcription challenge," 2020. [Online]. Available: http://catedrartve.unizar.es/reto2020/EvalPlan-S2T-2020-v1.pdf

[4] J. Jorge *et al.*, "MLLP-UPV and RWTH Aachen Spanish ASR Systems for the IberSpeech-RTVE 2018 Speech-to-Text Transcription Challenge," in *Proc. IberSPEECH 2018*, 2018, pp. 257–261.

[1] https://mllp.upv.es/git-pub/jjorge/MLLP_Streaming_API

[2] https://ttp.mllp.upv.es/

[5] E. Lleida *et al.*, "RTVE2018 database description," 2018. [Online]. Available: http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf

[6] E. L. et al., "Albayzin 2018 Evaluation: The IberSpeech-RTVE Challenge on Speech Technologies for Spanish Broadcast Media," *Applied Sciences*, vol. 9, no. 24, p. 5412, 2019.

[7] E. Lleida *et al.*, "RTVE2020 database description," 2020. [Online]. Available: http://catedrartve.unizar.es/reto2020/RTVE2020DB.pdf

[8] "Voxforge." [Online]. Available: http://www.voxforge.org

[9] M. del Agua *et al.*, "The translectures-UPV toolkit," in *Advances in Speech and Language Technologies for Iberian Languages*, Nov. 2014, pp. 269–278.

[10] S. J. Young *et al.*, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. of Workshop on Human Language Technology*, 1994, pp. 307–312.

[11] A. Zeyer *et al.*, "A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition," in *Proc. of ICASSP*, 2017, pp. 2462–2466.

[12] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015.

[13] D. S. Park *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. of Interspeech*, 2019, pp. 2613–2617.

[14] J. Jorge *et al.*, "Real-Time One-Pass Decoder for Speech Recognition Using LSTM Language Models," in *Proc. of Interspeech*, 2019, pp. 3820–3824.

[15] A. Stolcke, "SRILM - an extensible language modeling toolkit." in *Proc. of Interspeech*, 2002, pp. 901–904.

[16] Y. Lin *et al.*, "Syntactic annotations for the google books ngram corpus," in *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, 2012.

[17] X. Chen *et al.*, "CUED-RNNLM – An open-source toolkit for efficient training and evaluation of recurrent neural network language models," in *Proc. of ICASSP*, 2016, pp. 6000–6004.

[18] A. Mnih and Y. W. Teh, "A fast and simple algorithm for training neural probabilistic language models," *arXiv preprint arXiv:1206.6426*, 2012.

[19] X. Chen *et al.*, "Improving the training and evaluation efficiency of recurrent neural network language models," in *Proc. of ICASSP*, 2015, pp. 5401–5405.

[20] M. Ott *et al.*, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proc. of NAACL-HLT*, 2019, pp. 48–53.

[21] P. Baquero-Arnal *et al.*, "Improved Hybrid Streaming ASR with Transformer Language Models," in *Proc. of InterSpeech*, 2020, pp. 2127–2131.

[22] "OpenSubtitles," http://www.opensubtitles.org/.

[23] "UFAL Medical Corpus," http://ufal.mff.cuni.cz/ufal_medical_corpus.

[24] "Wikipedia," https://www.wikipedia.org/.

[25] C. Callison-Burch *et al.*, "Findings of the 2012 workshop on statistical machine translation," in *Proc. of WMT*, 2012, pp. 10–51.

[26] "News Crawl corpus (WMT workshop) 2015," http://www.statmt.org/wmt15/translation-task.html.

[27] "Eldiario.es," https://www.eldiario.es/.

[28] "ElPeriodico.com," https://www.elperiodico.com/.

[29] "CommonCrawl 2014," http://commoncrawl.org/.

[30] J. Jorge *et al.*, "LSTM-Based One-Pass Decoder for Low-Latency Streaming," in *Proc. of ICASSP*, 2020, pp. 7814–7818.

[31] J. Jorge *et al.*, "Live Streaming Speech Recognition using Deep Bidirectional LSTM Acoustic Models and Interpolated Language Models," submitted to: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Incorporation of a module for automatic prediction of oral productions quality in a learning video game

David Escudero-Mancebo, Valentín Cardeñoso-Payo, Mario Corrales-Astorgano, César González Ferreras, Valle Flóres-Lucas[1], Lourdes Aguilar[2], Yolanda Martín-de-San-Pablo[3], Alfonso Rodríguez-de-Rojas[4]

[1]Department of Computer Science, University of Valladolid, Spain
[2]Spanish Philology Department, Universitat Autònoma de Barcelona, Spain
[3]Fundación Personas, Valladolid, Spain
[4]Asociación Down Valladolid, Spain

descuder@infor.uva.es, valen@infor.uva.es

Abstract

This document presents the research project TIN2017-88858-C2-1-R of the Spanish Government for the incorporation of a module for automatic prediction of oral productions quality focusing on prosody in a learning video game adapted for Down syndrome speakers. It is our goal to present the project in Iberspeech 2020 as it is the main conference gathering specialists on speech technology in Iberian languages. We present the starting point of the project detailing antecedents of the learning game, the collection of human based evaluations during different testing campaigns of the software, the use of the collected data for training automatic assessment components; its integration in the video game for providing autonomous gaming; and the usability tests. As result, the new version of the video game permits autonomous and semi-supervised training thanks to the devised component for evaluating prosodic quality. The project concludes with relevant contributions to state of the art such as an annotated corpus, an open-source learning game and relevant analysis on the atypical prosodic patterns of Down syndrome speakers. [1]

Index Terms: Computer Assisted Pronunciation Training, Prosody, Down syndrome speech, Learning video games.

1. Introduction

This project continues the research line opened in 2014 with the project *La piedra mágica*[2] supported by Resercaixa and continued in 2016 with the project *Pradia* [3] supported by BBVA Humanidades Digitales. The first of these projects served to develop a video game for Down syndrome adolescents to train oral communication, in particular prosody and pragmatics [1]. In the second project, the video game routines were enriched and a clinical evaluation of user's performance was tested with formal evaluation.

The video game has the structure of a graphic adventure game, including conversations with characters and navigating through scenarios [1]. It includes three types of activities: comprehension, production and visual cognitive activities. Comprehension activities are focused on lexical-semantic comprehension and the improvement of prosodic perception in specific contexts. In these activities, players have to choose between different options to continue a conversation with a game character. Production activities are focused on oral production, so the player is encouraged by the game to train his/her speech, keeping in mind such prosodic aspects as intonation, expression of emotions or syllabic emphasis. In these activities, the player is introduced by the game to different conversations with game characters, where the player has to choose between different options to continue the dialogue or to record some sentences related with the dialogue context, depending on the activity. Finally, the visual activities are introduced to add variety to the game and to practice other skills not related directly with language. All activities have a maximum of attempts to avoid frustration in the players, but this maximum depends on the activity.

First we state the goals of the project as they appear in the original application form, next we report on different relevant project management concerns that has to be taken into account in the section of results which contents are aligned with the main publications derived from the project. We end with conclusions and future work.

In this context, in the first usability tests campaigns we detected that the demand of the game was high because it was an attractive resource both for students and for teachers. Nevertheless the original use of the game required the active participation of a therapist who assisted the intellectual disabled user while playing and decided whether he or she must or not repeat the oral production activity. The need of an specialist seated by the student during the training sessions limited the use of the system so that incorporating an automatic module that play the role of the therapist was a need. The training of such system and its incorporation in the learning game without degrading the quality of the training sessions was the challenge of the project.

2. Goals of the project

The video games developed in previous projects require the permanent assistance of a human trainer (a teacher, a therapist, a relative...) to control and monitor the progress of the game users. The goal of the current project is programing a module to increase the software capabilities so that users can play in an autonomous way supervised by an intelligent tutor. This intelligent system will be responsible to decide on the user to repeat or continue with the training activities giving feedback for him/her to be more competent in oral communication and prosodic skills.

Three stages have been defined: 1) corpus collection of audio of Down syndrome (DS) people playing with the video

[1]We would like to thank Ministerio de Economía y Competitividad y Fondos FEDER project key: TIN2017-88858-C2-1-R

[2]http://prado.uab.es/recercaixa/

[3]http://www.pradia.net

10.21437/IberSPEECH.2021-26

game; the DS audios will be rated by professional voice therapists; correcting advices of the therapists will be also monitored. 2) computational models of quality of DS turns will be trained from the audio; a knowledge data base will be compiled with the therapist corrective orders. 3) An expert system is using the compiled information to decide about the user activities in real time.

3. Project management

The IPs of the project (first two authors of this paper) have been responsible of the project management. The participation of the different members of the project is reflected in the list of authors of the different publications related with the achievements. Special mention must be done about the role of Lourdes Aguilar who signed a project proposal originally coordinated with the present one. Her proposal was rejected but has given support in what concerns to aspects related to linguistics, prosody and pragmatics. We had the opportunity to hire Mario Corrales whose technical and research initiative has been crucial in the development of the project.

We have collaborated with the research group of Pastora Martínez from the Department of Psychology of UNED Madrid. She is co-author of several relevant publications in the field of Down syndrome and prosody [2, 3]. She has provided us with a paired corpus of typical-DS speech, obtained while applying the PEPS-C test [4].

As stakeholders we counted with Fundación Personas Valladolid and with ASDOVA (Down syndrome Valladolid Association). Therapists and teachers of these institutions (in the list of coauthors) collaborated in the definition of the evaluation template, performed the evaluation and selected the informants. They also actively participated in the usability tests.

The Ethics Committee assistance and consent was required and one of the partners informed about changes in the regulation with respect to image rights of the informants. We specially thanks the support of Ricard Martínez on the compilation of the informed consents. Finally we receive positive authorization PI 20-1639.

The lock-down due to the pandemic COVID-19 interrupted our evaluation sessions in the ASDOVA center. Both the collection of the evaluation template sessions and the usability test campaign had to be postponed. Finally we could perform them during october/november 2020 but we had to ask for a 6 months prorogation so that the project is still alive until June 2021.

4. Achievements

Different parts of this sections are aligned with different articles written during the project development. Some of them are already published, like the ones presented in sections 4.1.1, 4.1.2 and 4.3.2; others are under revision like the ones presented in sections 4.1.3 or 4.3.1; and the one presented in 4.2 is in process.

4.1. Evaluation of prosodic quality of Down syndrome speakers

The works presented in this section are all of them related with the analysis of the prosodic features with different goals. Subsection 4.1.1 presents a work in which the acoustic prosodic features of Down syndrome speakers are compared with the ones of typical development speakers; the whole work can be found in [5]. Subsection 4.1.2 presents an analysis of the quality of the oral production of Down syndrome speakers from the prosodic acoustic features; the whole work can be found in [6] and in [7]. Finally, 4.1.3 presents an analysis of the prosodic patterns frequently observed in Down syndrome speech when different prosodic functions are performed; further details can be found in [8].

4.1.1. Differences between Down syndrome and typical development speakers

There are many studies that identify important deficits in the voice production of people with Down syndrome. These deficits affect not only the spectral domain, but also the intonation, accent, rhythm and speech rate. The main aim of this work was the identification of the acoustic features that characterize the speech of people with Down syndrome, taking into account the different frequency, energy, temporal and spectral domains. The comparison of the relative weight of these features for the characterization of Down syndrome people's speech was another aim of this study.

The openSmile toolkit with the GeMAPS feature set was used to extract acoustic features from a speech corpus of utterances from typically developing individuals and individuals with Down syndrome. Then, the most discriminant features were identified using statistical tests. Moreover, three binary classifiers were trained using these features.

The best classification rate, using only spectral features, is 87.33%, and using frequency, energy and temporal features, it is 91.83%. Finally, a perception test was performed using recordings created with a prosody transfer algorithm: the prosody of utterances from one group of speakers was transferred to utterances of another group.

The results of this test show the importance of intonation and rhythm in the identification of a voice as non typical. As conclusion, the results obtained point to the training of prosody in order to improve the quality of the speech production of those with Down syndrome.

4.1.2. Analysis of the impact of the speaker in the evaluation

Prosodic skills are useful for improving the communication of individuals with intellectual and developmental disabilities. Yet, the development of technological resources that consider these skills has received little attention. One reason that explains this gap is the difficulty of including an automatic assessment of prosody that considers the high number of variables and the heterogeneity of such individuals.

In this work, we analysed how the heterogeneity of people with Down syndrome can affect the automatic assessment of prosodic quality. To do this, a therapist and an expert in prosody judged the prosodic appropriateness of individuals with Down syndrome speech samples collected with a video game. The judgments of the expert were used to train an automatic classifier that predicts the quality by using acoustic information extracted from the corpus, with a classification rate of 79.3%.

In addition, the relationship of some prosodic features with the expert assessment of five speakers was also analyzed. We observe the different importance of the prosodic features in the automatic classification of the recordings of each speaker.

This result seems to indicate that this heterogeneity must be taken into account when developing an automatic assessment of the prosodic quality of people with Down syndrome.

Figure 1: *Evaluation sesion with real users.*

4.1.3. Analysis of the impact of the prosodic function

The speech of people with Down syndrome shows prosodic features which are distinct from the ones observed in the oral productions of typically developing speakers. Although a different prosodic realization does not necessarily imply wrong expression of prosodic functions, atypical expression may hinder communication skills. To ascertain whether this can be the case in individuals with DS is the focus of this work.

We analyzed the acoustic features that better characterize utterances of speakers with Down syndrome when expressing prosodic functions related to emotion, turn-end and phrasal chunking, and compare them with those used by typical development speakers. An oral corpus of speech utterances has been recorded using the PEPS-C prosodic competence evaluation tool.

We used automatic classifiers to prove that the prosodic features that better predict prosodic functions in typical development speakers are less informative in speakers with DS.

Although atypical features are observed in speakers with DS when producing prosodic functions, the intended prosodic function can be identified by listeners and, in most cases, the features correctly discriminate the function with analytical methods. However, a greater difference between the minimal pairs presented in the PEPS-C test is found for typical development speakers in comparison with DS speakers.

The proposed methodological approach provides, on the one hand, the identification of the set of features that distinguish the prosodic productions of Down syndrome and typical development speakers and, on the other, the set of target features for therapy of speakers with DS, from an analysis of the separation of prosodic functions.

4.2. Incorporation of the automatic module in the video game and usability tests

The use of information technologies is broadly extended among the population with intellectual disabilities, also, in lower degree, the use of tools for learning, including learning games. In spite of the use of learning games is widely accepted by the community, due to its great engaging capabilities, there are few works showing its efficiency.

In this work we present the study of the efficiency of the developed learning game by comparing the usability of the system in training sessions with different degrees of supervision of the player by the therapist.

We use the PRADIA software, the learning video game for the training of oral productions (prosody and pragmatics) that showed high user satisfaction rates in previous works [1] when is was used with the assistance of a therapist. We included a module that provides automatic evaluation of the players oral productions allowing autonomous use of the tool. The use of the game was compared in three scenarios: supervised, autonomous and semi-autonomous (groups of students playing in parallel with the assistance of a teacher).

Different instruments of usability evaluation reveal that, in spite of there are no differences in the degree of engaging, there could be differences in the profiting of the training sessions: lower quality of the recorded audios and more errors in the more autonomous playing modes.

We conclude that Down syndrome people autonomous training is possible (the main goal of the project), implying saving human resources costs, but the performance is highly dependent on the feedback provided. In spite of the degree of engagement of the autonomous version is high, the quality and quantity of feedback is not comparable with the one provided by the therapists resulting on lower performance in the training sessions.

4.3. Resources

The works presented in this section have to do with the resources generated during the project. They are freely available for the research community and they consist of an annotated corpus of Down syndrome utterances and the video game itself. Subsection 4.3.1 presents the corpus (further details will be found in [9]); and subsection 4.3.2 the video game with detais in [10] and [11].

4.3.1. The corpus of Down syndrome recordings

Oral productions of speakers with Down syndrome exhibit special characteristics that have been the target of study for decades. In spite of this attention, the availability of rich resources for its analysis is still scarce. In this project, we present the definition and compiling procedure of a corpus of semi-controlled oral productions of speakers with Down syndrome that aims to allow the analysis of how speakers with Down syndrome produce functional and linguistic aspects of speech.

The corpus (named PRAUTOCAL) has been recorded while speakers with Down syndrome use a video game for training oral competences. Utterances are related to well defined communicative tasks recorded by both speakers with Down syndrome and typically developing speakers. We present the procedure for human experts to evaluate the recordings and the transcription criteria followed for enriching the utterances of the corpus.

Although the activities of the video game mainly focus on prosody and pragmatics, we show that PRAUTOCAL permits the analysis of the clear contrast in voice and speech between individuals with Down syndrome and typically developing speakers, taking into account the high heterogeneity of the speech problems characteristic of the syndrome.

This material allows the analysis of the speech problems characteristic of the syndrome with applications to the generation of knowledge of the particular problem of these speakers that could be used in future works for therapists to prepare specific training or enriching diagnosis regarding possible speech and language disorders.

4.3.2. Disemination of the video game

With the growth in popularity of video games in our society many teachers have worked to incorporate gaming into their classroom. It is generally agreed that by adding something fun to the learning process students become more engaged and, consequently, retain more knowledge. However, although the characteristics of video games facilitate the dynamics of the educational process it is necessary to plan a pedagogical project that includes delimitation of learning goals and profile of the addressees, the conditions of application of the educational project, and the methodologies of evaluation of the learning progress.

This is how we can make a real difference between gamification and video game based learning. The paper addresses the design of an educational resource for special education needs students that aims to help teach communicative skills related to prosody. The technological choices made to support the pedagogic issues that underlie the educational product, the strategies to convert learning content into playful material, and the methodology to obtain measures of its playability and effectiveness are described.

The results of the motivation test certified that the video game is useful in encouraging the users to exercise their voice and the indicators of the degree of achievement of the learning goals serve to identify the most affected prosodic skills.

5. Conclusions and future work

We conclude that the goal of the the project has been fulfilled with the definition of a new module of the video game that permit autonomous gaming to Down syndrome players. The module for the evaluation of quality of learners oral productions has shown to be effective deciding whether players can continue playing or have to repeat the activities. Additionally, it has been done without degrading the engagement of users in the video game.

This project contribute to the building of knowledge about Down syndrome speech and voice, with the identification of acoustic prosodic features that better characterize it; we have identified atypical prosodic patterns used by this type of speakers and we have devised strategies to face up this peculiarities in the challenging frame of automatic assessment of prosody.

We are still working on the improvement of module for quality assessment. In particular, until the end of the project (June 2021) we are using the data collected in the usability tests to test and train a classifier that is specialized in the particular activity that is presented by the video game. The learning video game presents different activities for training different prosodic and pragmatic functions. Experimental results [12] show that they must be taken into account for being efficient on the evaluation of prosodic quality.

It has been applied a new research project to the Spanish Ministry for continuing this research line. In particular, we want to deepen into user adaptation concerns. Experimental results have shown that evaluation of prosodic quality highly depends on the particular speaker that is under analysis as DS speakers present a broad spectrum of not only speech deficits but also short memory limitations and intellectual disability. Our proposal is using the game records of the user for categorizing him/her and adapting the game experience in function of the game profile. Important benefits in terms of usability are expected to be obtained with this promising approach.

6. Acknowledgments

Special thanks to Jesús Gómez from CEIP Urueña for the collaboration during the evaluation template definition and during the inter-rater consistency tests.

7. References

[1] C. González-Ferreras, D. Escudero-Mancebo, M. Corrales-Astorgano, L. Aguilar-Cuevas, and V. Flores-Lucas, "Engaging adolescents with Down syndrome in an educational video game," *International Journal of Human–Computer Interaction*, vol. 33, no. 9, pp. 693–712, 2017.

[2] P. Martínez-Castilla and S. Peppé, "Developing a test of prosodic ability for speakers of iberian spanish," *Speech Communication*, vol. 50, no. 11-12, pp. 900–915, 2008.

[3] S. J. Peppé, P. Martínez-Castilla, M. Coene, I. Hesling, I. Moen, and F. Gibbon, "Assessing prosodic skills in five european languages: Cross-linguistic differences in typical and atypical populations," *International journal of speech-language pathology*, vol. 12, no. 1, pp. 1–7, 2010.

[4] S. Peppé and J. McCann, "Assessing intonation and prosody in children with atypical language development: the peps-c test and the revised version," *Clinical Linguistics & Phonetics*, vol. 17, no. 4-5, pp. 345–354, 2003.

[5] M. Corrales-Astorgano, D. Escudero-Mancebo, and C. González-Ferreras, "Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome," *Speech Communication*, vol. 99, pp. 90–100, 2018.

[6] M. Corrales-Astorgano, P. Martínez-Castilla, D. Escudero-Mancebo, L. Aguilar, C. González-Ferreras, and V. Cardeñoso-Payo, "Automatic assessment of prosodic quality in down syndrome: Analysis of the impact of speaker heterogeneity," *Applied Sciences*, vol. 9, no. 7, p. 1440, 2019.

[7] M. Corrales-Astorgano, P. Martínez-Castilla, D. Escudero-Mancebo, L. Aguilar, C. González-Ferreras, and V. Cardeñoso-Payo, "Towards an automatic evaluation of the prosody of people with Down syndrome," in *Proc. IberSPEECH 2018*, 2018, pp. 112–116. [Online]. Available: http://dx.doi.org/10.21437/IberSPEECH.2018-24

[8] M. Corrales-Astorgano, D. Escudero-Mancebo, C. González-Ferreras, V. C. Payo, and P. Martínez-Castilla, "Analysis of atypical prosodic patterns in the speech of people with down syndrome," *Biomedical Signal Processing and Control*, p. under evaluation, 2021.

[9] D. Escudero-Mancebo, M. Corrales-Astorgano, V. Cardeñoso-Payo, L. Aguilar, C. González-Ferreras, P. Martínez-Castilla, and V. Flores-Lucas, "Prautocal corpus:a corpus for the study of Down syndrome prosodic aspects," *Language Resources and Evaluation*, p. under evaluation, 2021.

[10] L. Aguilar, "Learning prosody in a video game-based learning approach," *Multimodal Technologies and Interaction*, vol. 3, no. 3, p. 51, 2019.

[11] F. Adell, L. Aguilar, M. Corrales-Astorgano, and D. Escudero-Mancebo, "Proceso de innovación educativa en educación especial: Enseñanza de la prosodia con fines comunicativos con el apoyo de un videojuego educativo," *Humanidades Digitales, Retos, Recursos y Nuevas Propuestas*, pp. 277–293, 2018.

[12] D. Escudero-Mancebo, M. Corrales-Astorgano, C. González-Ferreras, and V. Cardenoso-Payo, "Prosodic feature selection for automatic quality assessment of oral productions in people with Down syndrome," *Iberspeech 2020*, p. under revision, 2021.

CIRUSS Platform: Surgery Patient Empowerment by Stress and Anxiety Monitoring

Sergio Figueiras[1], *Alejandro García-Caballero*[2], *Carmen Garcia-Mateo*[3], *Laura Docio-Fernandez*[3], *Edward L. Campbell*[3], *Baltasar G. Perez-Schofield*[4], *Leandro Rodríguez-Liñares*[4], *Arturo J. Méndez*[4]

[1] Bahía Software [2] Fundación Biomédica Sur de Galicia [3] AtlanTTic - Universidade de Vigo - 36310 Vigo (Spain) [4] Department of Computer Science, ESEI, Universidade de Vigo

sergio.figueiras@bahiasoftware.es, alejandro.alberto.garcia.caballero@sergas.es, carmen.garcia@uvigo.es, ldocio@gts.uvigo.es, ecampbell@gts.uvigo.es, jbgarcia@uvigo.es, leandro@uvigo.es, mrarthur@uvigo.es

Abstract

In this paper, we present the CIRUSS software platform that has been developed within the framework of the Phase 2 of the pre-commercial procurement (PCP) for the Horizon 2020 STARS project. STARS-PCP aims at developing novel personalised solutions for reducing stress related to surgical procedures. CIRUSS is an integrated, scalable, sustainable and technologically adapted solution that aims to response to present and future needs of the European healthcare systems in relation to stress and anxiety management for surgery patients. Among the different functionalities included in CIRUSS, the solution integrates a tool for detecting the stress and anxiety by joint processing of voice, face and heart-rate of the patient. As a side product, a dataset of patient video interviews has been designed and acquired. This dataset has been used to assess the performance of monomodal stress detection systems as well the multimodal approach.

Index Terms: stress and anxiety detection, patient journey, voice analysis, machine learning

1. Introduction

Stress and anxiety are part of human existence. All people feel stress and anxiety in a moderate degree, as an adaptive response. It is important to understand stress as a feeling or an emotional state that help us to face stressful everyday situations. However, stress and anxiety in patients who must undergo surgery involves a significant negative emotional state, generating a physiological activation in the preparation of the organism to cope with the perceived danger, which can impair the correct development of the surgical procedure. As a result, stress causes an increase in postoperative pain, greater need for painkillers and prolongation in hospital stay days [1, 2]. All these factors have a direct impact on the cost of care and health system sustainability.

The Horizon 2020 funded project named STARS "Empowering Patients by Professional Stress Avoidance and Recovery Services" (https://stars-pcp.eu/) openly challenged the industry and research sectors to develop novel personalised solutions aimed to reduce stress related to surgical procedures. Reduction of the stress, experienced by patients, will lower the harmful side-effects of sedating drugs, shorten hospital stay, shorten recovery times and relieve carers and clinical staff from continuous assistance. STARS, whose consortium is composed of five leading European hospitals, started in January 2017 and will end in October 2021.

STARS Pre-commercial procurement (PCP) looked for the procurement of research and development of new innovative solutions before they are commercially available. STARS-PCP involved different suppliers competing through different phases of development (see Figure 1). Risks and benefits are shared between the procurers and the suppliers under market conditions.

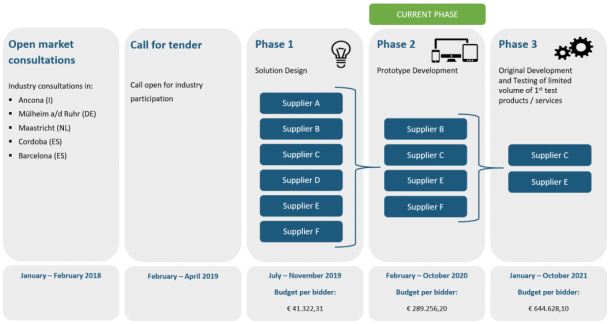


Figure 1: *STARS Pre-Commercial Procurement consists of three phases (extracted from https://stars-pcp.eu/pcp-phases)*

The Spanish company Bahía Software is one of the Phase 2 awarded suppliers to the STARS-PCP with a software solution named CIRUSS Platform (hereinafter, CIRUSS).

Next, the main components of CIRUSS are described, along with a preclinical study on automatic stress-anxiety detection.

2. Description of CIRUSS Platform

Although stress and anxiety can manifest in the different stages of the surgery patient journey, each patient will manifest stress in a more prominent way or another according to their basic psychology. There are never two identical situations or patients. Patients are extremely diverse, and so are the unique constellations of psychological problems people experience during the surgical process. Patients' symptoms of stress or anxiety may look very different from another patient's. Therefore, while certain types of stress reduction technologies may be deemed effective for some patients, they likely do not work equally well for all individuals, and that is why patients demand innovative interventions for stress reduction. Among the stress reassuring intervention strategies, the CIRUSS team has dedicated big efforts to the development of remote stress monitoring technologies. Access to high-quality stress monitoring data is compulsory to later take the right decisions about stress reassuring interventions.

10.21437/IberSPEECH.2021-27

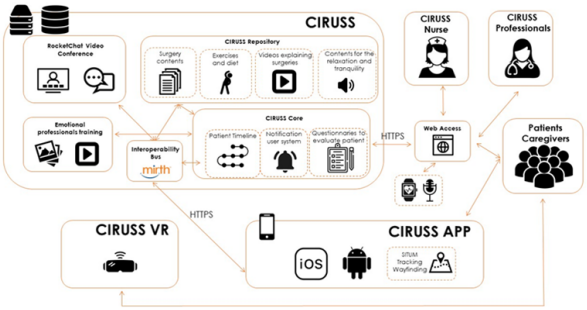


Figure 2: *Overall description of the CIRUSS platform.*

Figure 3: *Functional modules of the CIRUSS platform.*

Constant innovation on new assessing stress solutions in patients is of paramount importance. Real-time stress monitoring still represents today a major research challenge. There are different approaches to measure stress on patients, some of them more behavioural and others more biological. By these, we mean that some answers to the challenge try to understand how the individual is behaving, e.g., body language, while the latter approach relies more on measuring specific personal body parameters which can provide an indicator, e.g. high blood pressure as a potential indicator of stress.

Considering all the current challenges, our team has been working on the development of an integrated, scalable, sustainable and technologically adapted solution called the CIRUSS Platform (see Figure 2) that aims to better contribute to the stress and anxiety management of patients. We are aiming to build a flexible modular solution which

- can be easily integrated into existing ICT systems of a private, social and medical nature,
- can be tailored to the use and needs of the patient and other end-users, and
- incorporates a comprehensive set of intelligent functionalities for the assessment and management of stress in surgery patients.

The CIRUSS software platform intends to provide a comprehensive solution to cover the several gaps in surgery patient's journey. The added value offered to patients will start with

- providing valuable information to the patient, for patient empowerment and through the entire Patient Surgery journey. Firstly, CIRUSS will contain videos and materials to explain surgeries and patient surgery journeys. Secondly, CIRUSS will provide stress reassuring content (audios, videos, and materials based on Virtual Reality (VR)) to reduce stress in patients. Importantly, these materials will be based on stress reduction and online Cognitive Behavioral Therapies.
- evaluation of personalized patient stress patterns through wearables and advanced speech recognition technology.
- access to physical exercise perioperative programs and contents to promote exercise and good nutrition in surgery patients.
- indoor positioning technology to guide patients to specific points within hospitals and to inform caregivers about the position of their relative in the hospital in real time.
- videoconference services with the health professional designed by the Hospital. The videoconference service will be adapted to health professional preferences.

This platform consist of ten functional modules depicted in Figure 3. Some of the current challenges associated to stress and anxiety in surgery patients could be partially covered by them:

- improvement on the assessment of stress and anxiety in surgery patients.
- improving communication between the patient and health professionals by providing information appropriate to the patient's needs.
- reduction of cases of pre- and postoperative complications and adverse pathophysiological responses caused by stress.
- reduction of drug consumption and medications during the complete patient journey
- reduction of the time to return to daily routines.

3. Pre-clinical research study

As it was discussed before, the development of a system that automatically detects stress and anxiety is a challenge for the research community. Machine learning techniques could be applied if in-task data were available. That is why, in this project it was decided to conduct a preclinical research aiming at developing a non-intrusive technology that could serve to remotely assess stress in surgery patients. This study includes the design and acquisition of a dataset of video interviews with actual patients. Patient heart-rate was also recorded in order to be able to jointly process the voice, facial expression and heart rate of the patient.

This study has been conducted by researchers of the "Fundación Biomédica Sur de Galicia" and members of two different research groups of the University of Vigo. The data acquisition was performed in the Hospital of Ourense (Spain) and involved the recruitment of 65 patients under different surgery interventions (oncology, traumatology, etc). Currently, the dataset is being evaluated in order to determine stress levels (to collect samples after the surgery is also envisaged). This evaluation includes:

- the analysis of patient's stress levels by standard psychological instruments: Hospital Anxiety and Depression Scale (HADS), Ámsterdam Preoperative Anxiety and information Scale (APAIS) and Visual Analogue Anxiety Scale (VAS-A or EVA-A in Spanish)[3].

- classification of patients using the above mentioned psychological instruments in two groups: stressed and non-stressed using validated thresholds as a gold standard. Currently, the selection of the gold standard is still an open question.
- evaluation of patient's stress level using a salivary cortisol test [4].
- evaluation of heart rate variability(HRV) measures using the RHRV software package [5, 6, 7]: standard deviation of RR intervals (SDNN), inter-quartile (3rd and 1st) difference (IRRR), proportion of adjacent RR intervals differing by more than 50 ms (pNN50), standard deviation of differences between adjacent RR intervals (SDSD), square root of the mean of the squares of differences between adjacent RR intervals (rMSSD), median of the absolute differences between adjacent RR intervals (MADRR), HRV triangular index, calculated as the integral of the intervals histogram divided by its maximum (HRVi), approximate entropy (ApEn), and parameters SD1 and SD2 obtained from the Poincaré plot. SD1 is usually calculated as the standard deviation of the points perpendicular to the line of identity, and SD2 is calculated as the standard deviation along the line of identity.
- evaluation of patient's stress level by processing his/her speech. This analysis is based on iVector technology [8]. i-vector can be considered as a speech embedding that preserves the spectral patterns in the speaker's voice that allow the distinction between patient with and without stress and anxiety.
- evaluation of Ekman basic facial emotions using Sightcorp Emotion Recognition[1] software.The relevant categories for this project are Fear and Surprise.
- development of fusion strategies that combine the output of individual classification systems.

The purpose of this research was to explore the applicability in real clinical conditions of a standardized patient interview for stress evaluation including voice recognition, HRV and facial emotion recognition. After patient classification using international gold standards (HAD, APAIS and VAS-A) the sample was classified in two groups (stressed and non-stressed) and after normalization of the outputs of the different experimental measures, ROC curves were calculated for each measure and combining the best variables (Fusion). Fusion score showed an Accuracy of 0.777 and AUC (Area under the curve) of 0.800 (i.d. Fusion score (Fear, Surprise, MADRR, ApEn and Voice) classified correctly 80% of cases). After these promising results, challenge now is to explore the external validity of this approach to automatically and remotely evaluate psychological stress during the different stages of the patient's journey. The algorithms and mathematical models developed in Phase 2 of the STARS PCP project are now being integrated into the CIRUSS platform for this study.

4. Discussion and Future Work

CIRUSS Platform is being under evaluation by the STARS consortium. Overall, the preclinical research and the software developed (integrated into the CIRUSS platform), can offer significant novelty to the anxiety and stress monitoring field. Moreover, they can also provide recovery services through the entire surgery patient journey.

In case of being selected for Phase 3, CIRUSS will be verified and compared in terms of performance (interoperability, scalability, etc.) with other alternative solutions. This evaluation will be carried out in real-life operational conditions of the targeted public service.

[1] https://sightcorp.com/emotion-recognition/

5. References

[1] A. Rosiek, T. Kornatowski, A. Rosiek-Kryszewska, Ł. Leksowski, and K. Leksowski, "Evaluation of stress intensity and anxiety level in preoperative period of cardiac patients," *BioMed research international*, vol. 2016, 2016.

[2] L. Poole, A. Ronaldson, T. Kidd, E. Leigh, M. Jahangiri, and A. Steptoe, "Pre-surgical depression and anxiety and recovery following coronary artery bypass graft surgery," *Journal of Behavioral Medicine*, vol. 40, no. 2, pp. 249–258, 2017.

[3] E. Facco, G. Zanette, L. Favero, C. Bacci, S. Sivolella, F. Cavallin, and G. Manani, "Toward the validation of visual analogue scale for anxiety," *Anesthesia progress*, vol. 58, no. 1, pp. 8–13, 2011.

[4] M. Laudat, S. Cerdas, C. Fournier, D. Guiban, B. Guilhaume, and J. Luton, "Salivary cortisol measurement: a practical approach to assess pituitary-adrenal function," *The Journal of Clinical Endocrinology & Metabolism*, vol. 66, no. 2, pp. 343–348, 1988.

[5] C. A. G. Martínez, A. O. Quintana, X. A. Vila, M. J. Lado, L. Rodríguez-Liñares, J. M. R. Presedo, and A. J. Méndez, *Heart rate variability analysis with the R package RHRV*. Springer, 2017.

[6] L. Rodríguez-Liñares, A. J. Méndez, M. J. Lado, D. N. Olivieri, X. Vila, and I. Gómez-Conde, "An open source tool for heart rate variability spectral analysis," *Comput. Meth. Progr. Bio.*, vol. 103, no. 1, pp. 39–50, 2011.

[7] "RHRV: Heart Rate Variability Analysis of ECG Data," http://rhrv.r-forge.r-project.org, modified: Oct. 2019 (ver. 4.2.5), accessed: Apr. 2020.

[8] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Voice Restoration with Silent Speech Interfaces (ReSSInt)

Inma Hernáez Rioja[1], *Jose A. Gonzalez-Lopez*[2], *Eva Navas*[1], *Jose Luis Pérez Córdoba*[2], *Ibon Saratxaga*[1], *Gonzalo Olivares* [3], *Jon Sanchez*[1], *Alberto Galdón*[3], *Victor García Romillo*[1], *Míriam González-Atienza*[2], *Tanja Schultz*[4], *Phil D. Green*[5], *Michael Wand*[6], *Ricard Marxer*[7], *Lorenz Diener*[4]

HiTZ Center - Aholab, University of the Basque Country UPV/EHU, Spain
[2]SigMAT group, University of Granada, Spain
[3]Hospital Universitario Virgen de las Nieves, Granada, Spain
[4]Cognitive Systems Lab, University of Bremen, Bremen, Germany
[5]Speech and Hearing Group, University of Sheffield, UK
[6]Swiss AI Lab, ISDIA, Manno, Switzerland
[7]University of Toulon, France

inma.hernaez@ehu.eus, joseangl@ugr.es, eva.navas@ehu.eus

Abstract

ReSSInt aims at investigating the use of silent speech interfaces (SSIs) for restoring communication to individuals who have been deprived of the ability to speak. SSIs are devices which capture non-acoustic biosignals generated during the speech production process and use them to predict the intended message. Two are the biosignals that will be investigated in this project: electromyography (EMG) signals representing electrical activity driving the facial muscles and invasive electroencephalography (iEEG) neural signals captured by means of invasive electrodes implanted on the brain. From the whole spectrum of speech disorders which may affect a person's voice, ReSSInt will address two particular conditions: (i) voice loss after total laryngectomy and (ii) neurodegenerative diseases and other traumatic injuries which may leave an individual paralyzed and, eventually, unable to speak. To make this technology truly beneficial for these persons, this project aims at generating intelligible speech of reasonable quality. This will be tackled by recording large databases and the use of state-of-the-art generative deep learning techniques. Finally, different voice rehabilitation scenarios are foreseen within the project, which will lead to innovative research solutions for SSIs and a real impact on society by improving the life of people with speech impediments.

Index Terms: Silent speech interfaces, brain to speech conversion, EMG to speech, speech synthesis, voice conversion, deep neural networks.

1. Introduction

Speech is the first and foremost means of human communication. Unfortunately, many people are not able to speak, in particular those who have lost this ability through illness or disability. There are no many studies providing specific data about the prevalence of this disability. In [1] the authors conclude that 0.4% of the European population suffer from a speech impediment. In a later survey from 2011 [2], it is reported that 0.5% of people in Europe present 'difficulties' with communication. Focusing on Spain (data from the Spanish National Institute for Statistics (INE) published in 2008) there are more than 410,000 people with a disability to produce spoken messages [3]. For instance, laryngectomy patients (∼1200 total laryngectomies are performed every year in Spain [4]), whose voice box has been completely removed to treat larynx cancer, can no longer speak in a conventional way after the operation. Speech is also affected after brain damage, spinal cord injuries or neurodegenerative diseases such as amyotrophic lateral sclerosis (ALS), a disease which is expected to increase worldwide by 69% between 2015 and 2040 [5] due to an aging population and improved public healthcare. As this disease progresses, individuals can no longer communicate verbally and assistive devices that rely on nonverbal signals are needed for communication.

Voice loss is not only a problem for efficient communication, but also deprives the speaker of a central personal characteristic (namely, his/her own voice) which can in turn lead to occupational disability, personal isolation, and clinical depression [6]. In the absence of clinical procedures for repairing the damage caused by the above disorders, several methods are used to restore communication. However, all traditional methods are, in general, far from ideal. For laryngectomized patients, the 'gold-standard' method for voice restoration, the tracheoesophageal valve, requires frequent replacement every 3-4 months due to biofilm growth [7] and produces a masculine voice disliked by female patients. The electrolarynx, on the other hand, despite being relatively easy to use and safe, produces a very robotic and monotone voice. Esophageal speech, a method of speech production that involves oscillation of the esophagus, sounds gruff and masculine and is difficult to master. Additionally, esophageal speech is less intelligible both by humans and machines ([8,9]), which makes voice interaction with computer very difficult. Although voice conversion strategies have been investigated to improve the quality and intelligibility of these voices, there is still margin for improvement [10, 11].

Things are even worse for people who have suffered a brain stroke or neurodegenerative disease. These patients normally find themselves struggling with communication and need to use alternative methods, such as augmentative and alternative communication (AAC) devices, to communicate with their family and caregivers. These devices, usually with rates lower than 15 words per minute, are only suitable for short conversations.

In recent years, SSIs [12–14] have emerged as a promising alternative to restore oral communication by decoding speech from non-acoustic (silent) speech-related biosignals generated during speech production. Lip reading is the best-known form of silent speech communication. A variety of sensing modalities have been investigated to capture those biosignals, such as

10.21437/IberSPEECH.2021-28

vocal tract imaging [15], electromagnetic articulography (magnetic tracing of articulator movements) [16, 17], EMG [18–20], which captures facial muscle activity using surface electrodes, and iEEG [21–23], which captures the electrical activity in the brain. Since SSIs allow to capture speech without requiring any acoustic signal at all, they offer a fundamentally new solution to restore communication capabilities to speech-disabled persons.

The project described in this paper will investigate SSIs-based communication systems for restoring vocal communication to individuals who have been deprived of the ability to speak. In particular, ReSSInt will address two user groups: total-laryngectomy patients and individuals affected by brain damages. Each speech impairment will be addressed by a specific interface: EMG and iEEG. In the following sections we present the main characteristics and relevant previous works using these interfaces. We also describe the main objectives of the project and how it has been structured to achieve the objectives.

2. Silent Speech Interfaces

Two approaches have been proposed to decode speech from silent, speech-related biosignals [14]: silent speech-to-text and direct speech synthesis. In the first approach, automatic speech recognition (ASR) algorithms trained on silent speech data are used to decode speech from the input data. Text-to-speech (TTS) software can then be used to synthesize speech from the decoded text if required. Several works have shown promising results on silent speech-to-text for a variety of methods. For instance, in [24], a EMG-based silent speech recognizer was proposed. The system was evaluated on a corpus of phonetically-rich sentences recorded by $n = 8$ healthy persons, achieving a word error rate (WER) of 16.8%. In [25], EMG-based silent speech recognition was evaluated as assistive communication device for $n = 8$ laryngectomized patients, achieving an average WER of 10.3%.

Although still in a more preliminary stage, speech decoding from recordings of neural activity in anatomical regions involved in continuous speech production has also been shown to be feasible. Due to the potential advantages of the brain-to-text approach, breakthrough advances have been achieved in recent years. Thus, [26] demonstrated that phonetic features, such as place and manner of articulation, and voicing status, can be decoded during continuous speech production from eletrocorticography. Mugler *et al.* [27] was the first study reporting on decoding the full set of phonemes for American English, obtaining up to 36% accuracy when classifying phonemes within word productions and up to 63% accuracy for a single phoneme. The first study to address the task of decoding continuous speech from iEEG recordings was [21]. Seven patients undergoing surgery for epilepsy treatment where implanted with electrocorticography (ECoG) sensors (a type of intracranial EEG technique where electrode strips are places directly over the exposed surface of the brain) and, later, speech and ECoG signals were simultaneously recorded while the subjects read texts aloud. Acquired brain signals were used to train speech recognizers for each subject. Up to 75% word accuracy was reported when the vocabulary consisted on 10 possible words, and up to 40% when the user could choose between 100 words.

2.1. Direct speech synthesis from silent speech data

Though appealing, the silent speech-to-text approach lacks the real-time capabilities (i.e. low latency) that a SSI system for natural human speech communication would require. In this regard, previous studies have provided estimates on the maximum latency for an ideal SSI system. In oral communication, 100 to 300 ms of propagation delay causes slight hesitation on a partner's response and beyond 300 ms causes users to begin to back off to avoid interruption [28]. Studies on delayed auditory feedback, in which subjects received delayed feedback of her/his voice, found disruptive effects on speech production in subjects with delays starting at 50 ms and maxing out around 200 ms [29]. Altogether, these results suggest a $\sim$50 ms latency for an ideal SSI system, though latencies up to $\sim$100 ms may still be reasonable. These low latencies can only be achieved through the second SSII approach, direct speech synthesis, in which audible speech is directly generated from silent speech data by mapping the input silent data into a suitable speech representation (e.g. MFCCs) and then generating a waveform from the estimated speech parameters. Most commonly, deep neural networks (DNNs) [30] trained on time-aligned speech and silent data recordings (i.e. parallel data) are used to model the silent speech-to-speech mapping.

The research team of ReSSInt, as well as our international collaborators, have made significant contributions on the direct speech synthesis approach. Thus, in [16], we proposed a SSI system based on direct speech synthesis and electromagnetic articulography. The vocabulary consisted on digit sequences and consonant-vowel pairs. Our results showed that intelligible speech could be generated from articulatory data, although some phones were more mistakable than others (i.e. phones differing in their manner of articulation or voicing were hard to distinguish from the silent speech data). Building upon this work, in [17], we addressed the task of synthesizing continuous speech for a large vocabulary using recurrent neural networks. On average, the resulting speech was $\sim$75% intelligible, but for some subjects speech intelligibility reached up to $\sim$92%.

Although direct speech synthesis from EMG signals has experienced considerable advances in recent years, this technology still presents many challenges which keep it from becoming a product. The first limitation is the strong dependency of the results on the training session. Although array EMG sensors have provided greater signal stability and robustness in this sense (thus the relative position of the sensors is kept constant), there are still differences in data between different sessions [31, 32]. In [33] the authors show that even the training material (style, isolated words, syllables) can influence the quality of the results. More importantly, all the mentioned experiments are carried out on a speaker dependent fashion and speaker independence remains unsolved. Finally, for a real application of this kind of systems, real time performance must be achieved. Although there have been some recent attempts [34], this is still an open research issue.

In parallel with these findings, direct synthesis from neural signals is gaining growing attention due to the promises of restoring speech function in individuals unable to speak. In comparison with other sensing techniques, recording silent speech data directly from the brain has the advantage that speech is captured in an earlier form, which means that audio could be synthesized from the neural signals with lower latency. To date, however, only a few studies have addressed the task of generating speech directly from neural activity. The first study to report on this was [35] in which neural activity recorded from a completely-paralyzed individual was used to decode formant frequencies during imagined speech. As the user became engaged in more training sessions, he quickly improved with practice, learning to control the system to produce

better acoustics. Martin *et al.* [36] investigated the prediction of continuous speech from ECoG during imagined speech. In this study, $n = 7$ subjects were asked to read aloud (overt speech) and imagined reading (covert speech) short paragraphs of text. Later, models were trained for the overt condition to predict speech parameters from ECoG. These models were also applied to predict the speech parameters during the covert condition. Despite not being fully intelligible, human listeners were able to identify the reconstructed speech when they have to choose from a list of sentences. More recently, [22] proposed a two-stage approach in which they first transformed ECoG signals into anatomical representations of the vocal-tract articulators, and then transformed such intermediate representations into speech using bidirectional recurrent neural networks. Reconstructed speech waveforms for $n = 5$ volunteers were deemed quite intelligible by human listeners. Our collaborators from the University of Bremen have also made important contributions in this field. For instance, in [23, 37], they investigated the synthesis of speech from ECoG signals. Two approaches were used to map ECoG into speech: 3-dimensional convolutional neural networks (CNNs) and a concatenative, unit-selection approach. In general, despite not being fully intelligible, it was found that synthesized speech sounded natural and included features such as prosody and accentuation. Moreover, it was found that speech motor cortex provided more information for the reconstruction process than the other cortical areas.

3. Objectives of the project

Despite the promising results and advances achieved so far, SSI devices have not made it to the mass market. In our opinion, a major reason for this is the lack of focus on real-life use cases. In particular, the problem of inter-session and inter-speaker variability is not yet solved and requires intensive further investigation. Also, most works have not taken full advantage of recent advances in generative DNNs. For example, most systems have focused on predicting the spectral envelope while nowadays neural vocoders can generate prosodic information as well. Furthermore, most existing studies have been performed with able-bodied subjects, often relying on parallel recorded silent speech-and-acoustic signals. This excludes the important group of speech-disabled persons who have already lost their voice or have it severely impaired. Finally, many studies have used offline data, which disregards the fact that a user will expect acoustic feedback during the process of speaking silently. This feedback will allow the user to improve/adapt his/her own speaking patterns (we speak of coadaptation of the user and the device). Additionally, care needs to be taken to make the system flexible and easy-to-use, which implies lightweight and portable devices, fast enrollment, and graceful degradation in the case of processing errors.

In ReSSInt, we intend to overcome the limitations of both traditional voice rehabilitation methods and previous SSI studies by investigating SSI-based communication systems for restoring communication to individuals who have been deprived of the ability to speak. From the whole spectrum of speech disorders which may affect a person's voice, ReSSInt will address two conditions, each being the objective of a particular subproject:

- **Subproject 1**: total-laryngectomy patients. These persons still retain the control over their speech articulators and, hence, silent speech data reflecting the movements of the articulators can be easily captured using EMG.
- **Subproject 2**: neurodegenerative diseases and other traumatic injuries which may leave an individual paralyzed and, eventually, unable to speak. For many of these individuals, the only means of communication is through limited eye movements and blinking; however, for those with complete paralysis, even this type of communication may not even be possible. An SSI-based communication system could provide a more effective and efficient way to communicate. Such a technology could dramatically improve these people's lives and, arguably, its potential benefits would outweigh the risks of brain surgery for implanting iEEG electrodes.

For an SSI system to be truly beneficial for these persons, it must satisfy the following criteria, which have guided us in the definition of the main goals of the project:

- It must be able to generate intelligible speech with a reasonable quality and naturalness.
- The SSI system needs to be robust to intra- and inter-speaker differences.
- The system must be flexible enough to deal with a variety of rehabilitation scenarios, in particular:
    1. Patients who are able to record synchronous silent speech and acoustic data before losing their voice,
    2. Recordings of a patient's original voice may be available, but silent speech biosignals is only recorded after s/he has completely lost her/his voice,
    3. No recordings of the original voice are available, so a substitute voice (e.g. a voice donor, perhaps a close relative) needs to be used instead. The third scenario is particularly relevant to SP2 given the difficulty of recording speech for paralyzed patients.
- Finally, a practical SSI must be able to generate audio from silent speech data in close to real-time (latency $<$100 ms), so its user can receive synchronous acoustic feedback while speaking and can adapt her/his articulation style to improve the output.

We will accomplish these goals by taking advantage of the background work on speech synthesis and SSIs of both groups and by recording large datasets, which in turn will foster the use of cutting edge deep learning techniques to improve the performance beyond the state-of-the-art. The real-time system will play a central role during the evaluation phase to assess the performance of the SSI in terms of speech intelligibility, quality, and naturalness. This system will also pave the way for studies of user-in-the-loop strategies, where both the user and the system co-adapt themselves to optimize the output.

Summarizing, the specific objectives of the coordinated project are:

1. To explore the paths and advances in the application of state-of-the-art deep generative neural network architectures to improve the present quality and intelligibility of current SSIs using EMG and ECoG.
2. To develop corpus, databases, protocols and best practices for research on SSI in Spanish language.
3. To establish a new research line and, consequently, a research infrastructure for SSI in Spain.
4. To strengthen the links between two of the most consolidated research groups on speech technologies at the national level: Aholab at UPV/EHU and SiGMAT at UGR.

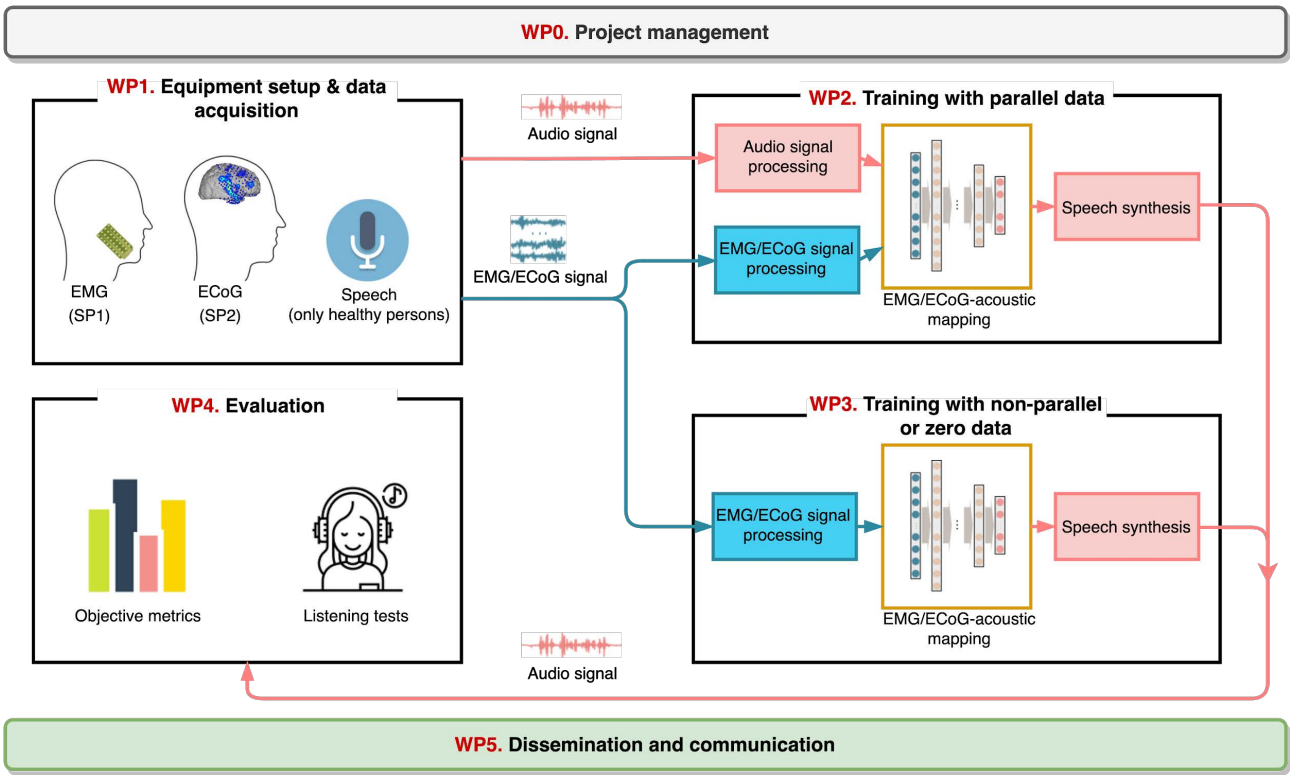


Figure 1: *Work package diagram for ReSSInt.*

4. Methodology

ReSSInt is split into 5 work packages, representing functional units which are necessary to tackle the goals of the project. The major workflow between the work packages is shown in Figure 1.

WP0 and WP5 are transverse work packages dedicated to the project management and dissemination/communication activities, respectively. WP1 is dedicated to the experimental part of the project (i.e. equipment setup, stimuli definition, participant recruitment and data acquisition). Data recorded in WP1 will be used by WP 2 and 3, where the foundational algorithms for direct speech synthesis from silent speech data will be developed. In particular, WP2 deals with training DNN architectures in a supervised fashion using parallel data. WP3, in contrast, is dedicated to training with non-parallel data. WP4 is dedicated to the evaluation of the algorithms and the user tests, which will provide continuous input, assessment, and improvement requests to the technical work packages. Our development model is iterative, with frequent interactions between work packages and partners. Most work packages run for the entire length of the project; yet research is structured by the subordinate tasks within the work packages.

5. Conclusions

In this paper we have presented the project ReSSInt, which will be executed in the period from July 2020 till June 2023. The project involves two research groups located in Spain (at the University of the Basque Country UPV/EHU and the University of Granada) in collaboration with expert researchers from other countries.

The beginning of the project has been greatly affected by COVID-19 and some of the task corresponding to the first year are suffering a delay. The acquisition of ECoG data in SP2 has not started yet, due to strict limitations on non-urgent surgery. Also, the acquisition of the equipment needed in SP1 has been delayed. However, the preparation of baseline systems is going on using external data provided by our collaborators. Our expectation is that we will recover from the initial delay and the main goals of the project remain viable.

Updated information about this project can be found at http://aholab.ehu.eus/ressint.

6. Acknowledgments

This work has been funded by the Agencia Estatal de Investigación ref.PID2019-108040RB-C21/AEI/10.13039/501100011033 and PID2019-108040RA-C22/AEI/10.13039/501100011033. Jose A. Gonzalez-Lopez holds a Juan de la Cierva-Incorporation Fellowship from the Spanish Ministry of Science, Innovation and Universities (IJCI-2017-32926).

7. References

[1] D. Dupré and A. Karjalainen, "Employment of disabled people in europe in 2002," *Statistics in focus*, pp. 3–26, 2003.

[2] Eurostat, "Population by type of basic activity difficulty, sex and age (hlth_dp040)," https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=hlth_dp040&lang=en, 2011, accessed: 23-10-2019.

[3] "Portal Web del Instituto Nacional de Estadística. Encuesta de Discapacidad, Autonomía Personal y Situaciones de Dependencia 2008," https://www.ine.es/jaxi/Tabla.htm?path=/t15/p418/a2008/hogares/p01/modulo1/l0/&file=01002.px&L=0, accessed: 2020-12-12.

[4] P. D. de Cerio Canduela, I. A. González, R. B. Durban, A. S. Suárez, M. T. Secall, P. L. P. Arias, C. of Head, L. R. W. Group *et al.*, "Rehabilitation of the laryngectomised patient. recommendations of the spanish society of otolaryngology and head and neck surgery," *Acta Otorrinolaringologica (English Edition)*, vol. 70, no. 3, pp. 169–174, 2019.

[5] K. C. Arthur, A. Calvo, T. R. Price, J. T. Geiger, A. Chio, and B. J. Traynor, "Projected increase in amyotrophic lateral sclerosis from 2015 to 2040," *Nature communications*, vol. 7, no. 1, pp. 1–6, 2016.

[6] H. Danker, D. Wollbrück, S. Singer, M. Fuchs, E. Brähler, and A. Meyer, "Social withdrawal after laryngectomy," *European Archives of Oto-Rhino-Laryngology*, vol. 267, no. 4, pp. 593–600, 2010.

[7] S. Ell, "Candida-the cancer of silastic," *Journal of laryngology and otology*, vol. 110, no. 3, pp. 240–242, 1996.

[8] S. Raman, I. Hernaez, E. Navas, and L. Serrano, "Listening to Laryngectomees: A study of Intelligibility and Self-reported Listening Effort of Spanish Oesophageal Speech," in *Proc. IberSPEECH 2018*, 2018, pp. 107–111. [Online]. Available: http://dx.doi.org/10.21437/IberSPEECH.2018-23

[9] S. Raman, L. Serrano, A. Winneke, E. Navas, and I. Hernaez, "Intelligibility and listening effort of spanish oesophageal speech," *Applied Sciences*, vol. 9, no. 16, p. 3233, 2019.

[10] L. Serrano, D. Tavarez, X. Sarasola, S. Raman, I. Saratxaga, E. Navas, and I. Hernaez, "Lstm based voice conversion for laryngectomees." in *IberSPEECH*, 2018, pp. 122–126.

[11] L. Serrano, S. Raman, D. Tavarez, E. Navas, and I. Hernaez, "Parallel vs. non-parallel voice conversion for esophageal speech." in *INTERSPEECH*, 2019, pp. 4549–4553.

[12] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[13] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257–2271, 2017.

[14] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. Martín-Doñas, J. L. Pérez-Córdoba, and A. M. Gomez, "Silent speech interfaces for speech restoration: A review," *IEEE Access*, vol. 8, pp. 177 995–178 021, 2020.

[15] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3444–3453.

[16] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "A silent speech system based on permanent magnet articulography and direct synthesis," *Computer Speech & Language*, vol. 39, pp. 67–87, 2016.

[17] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Direct speech reconstruction from articulatory sensor data by machine learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2362–2374, 2017.

[18] T. Schultz and M. Wand, "Modeling coarticulation in emg-based continuous speech recognition," *Speech Communication*, vol. 52, no. 4, pp. 341–353, 2010.

[19] M. Wand and T. Schultz, "Session-independent emg-based speech recognition." in *Biosignals*, 2011, pp. 295–300.

[20] L. Diener, M. Janke, and T. Schultz, "Direct conversion from facial myoelectric signals to speech using deep neural networks," in *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2015, pp. 1–7.

[21] C. Herff, D. Heger, A. De Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, "Brain-to-text: decoding spoken phrases from phone representations in the brain," *Frontiers in neuroscience*, vol. 9, p. 217, 2015.

[22] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[23] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz, "Speech synthesis from ecog using densely connected 3d convolutional neural networks," *Journal of neural engineering*, vol. 16, no. 3, 2019.

[24] M. Wand, M. Janke, and T. Schultz, "Tackling speaking mode varieties in emg-based speech recognition," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 10, pp. 2515–2526, 2014.

[25] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, "Silent speech recognition as an alternative communication device for persons with laryngectomy," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 12, pp. 2386–2398, 2017.

[26] F. Lotte, J. S. Brumberg, P. Brunner, A. Gunduz, A. L. Ritaccio, C. Guan, and G. Schalk, "Electrocorticographic representations of segmental features in continuous speech," *Frontiers in human neuroscience*, vol. 9, p. 97, 2015.

[27] E. M. Mugler, J. L. Patton, R. D. Flint, Z. A. Wright, S. U. Schuele, J. Rosenow, J. J. Shih, D. J. Krusienski, and M. W. Slutzky, "Direct classification of all american english phonemes using signals from functional speech motor cortex," *Journal of neural engineering*, vol. 11, no. 3, p. 035015, 2014.

[28] S. Na and S. Yoo, "Allowable propagation delay for voip calls of acceptable quality," in *International Workshop on Advanced Internet Services and Applications*. Springer, 2002, pp. 47–55.

[29] A. Stuart, J. Kalinowski, M. P. Rastatter, and K. Lynch, "Effect of delayed auditory feedback on normal speakers at two speech rates," *The Journal of the Acoustical Society of America*, vol. 111, no. 5, pp. 2237–2241, 2002.

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[31] M. Janke and L. Diener, "EMG-to-speech: Direct generation of speech from facial electromyographic signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2375–2385, 2017.

[32] L. Diener, G. Felsch, M. Angrick, and T. Schultz, "Session-independent array-based emg-to-speech conversion using convolutional neural networks," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[33] L. Diener, S. Bredehoeft, and T. Schultz, "A comparison of emg-to-speech conversion for isolated and continuous speech," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[34] L. Diener, C. Herff, M. Janke, and T. Schultz, "An initial investigation into the real-time conversion of facial surface emg signals to audible speech," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 888–891.

[35] F. H. Guenther, J. S. Brumberg, E. J. Wright, A. Nieto-Castanon, J. A. Tourville, M. Panko, R. Law, S. A. Siebert, J. L. Bartels, D. S. Andreasen *et al.*, "A wireless brain-machine interface for real-time speech synthesis," *PloS one*, vol. 4, no. 12, p. e8218, 2009.

[36] S. Martin, P. Brunner, C. Holdgraf, H.-J. Heinze, N. E. Crone, J. Rieger, G. Schalk, R. T. Knight, and B. N. Pasley, "Decoding spectrotemporal features of overt and covert speech from the human cortex," *Frontiers in neuroengineering*, vol. 7, p. 14, 2014.

[37] C. Herff, L. Diener, M. Angrick, E. Mugler, M. C. Tate, M. A. Goldrick, D. J. Krusienski, M. W. Slutzky, and T. Schultz, "Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices," *Frontiers in neuroscience*, vol. 13, p. 1267, 2019.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

The VoxSenes project: a study of segmental changes and rhythm variations on European Portuguese aging voice

Catarina Oliveira[1,2], *Ana Rita Valente*[1,3], *Luciana Albuquerque*[1,3,4,5], *Fábio Barros*[1,3], *Paula Martins*[1,2,6], *Samuel Silva*[1,3], *António Teixeira*[1,3]

[1]Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, Aveiro, Portugal
[2]School of Health Sciences, University of Aveiro, Portugal
[3]Dep. Electronics, Telecommunications and Informatics, University of Aveiro, Portugal
[4]Center for Health Technology and Services Research, University of Aveiro, Portugal
[5]Department of Education and Psychology, University of Aveiro, Aveiro, Portugal
[6]Institute of Biomedicine, University of Aveiro, Portugal

coliveira@ua.pt, rita.valente@ua.pt, lucianapereira@ua.pt, fabiodaniel@ua.pt, pmartins@ua.pt, sss@ua.pt, ajst@ua.pt

Abstract

The process of aging is generally associated with a number of changes in physiological, cognitive, psychological and social domains, including modifications on vocal quality of individuals. This paper presents a recent project - VoxSenes - that intends to bridge knowledge gaps in the speech changes due to aging. A deeper knowledge on how speech changes with age is essential for the development of automatic speech recognition systems suitable for older's voices and for clinical assessment of speech disorders. The VoxSenes project aims to study aging voice at segmental and suprasegmental level. The variation of acoustic parameters were analysed through audio speech samples and articulatory parameters were investigated over ultrasound tongue imaging. The most relevant age-related results of this project include: an increase in vowel duration, an approximation of F0 between genders, a centralization of the acoustic vowel space for males and an increase in speech pauses. The unsupervised method already developed to extract tongue contours from ultrasound tongue images provides the required data for an automatic analysis of relevant parameters to assess speech changes on vowel production.

An analysis of articulatory space of European Portuguese vowels is ongoing for speakers of different age groups, as well as a longitudinal study of age-related changes in speech rhythm.
Index Terms: Aging speech, Acoustic, European Portuguese, Speech production

1. Introduction

Population aging is one of the greatest triumphs of modern society, but also one of its major challenges [1]. More than in other developed countries, the Portuguese population has been aging (between 1970 and 2018, the percentage of people aged 65 and over increased from 9.7% to 21.8% [2, 3], both because of increased life expectancy and the decrease in fertility rates [1].

The process of aging is generally associated with physiological, cognitive, psychological and social changes, including modifications on individuals' vocal parameters. From childhood to old age, the speech production mechanism undergoes numerous anatomical and physiological changes in the respiratory, laryngeal and supralaryngeal system [4, 5, 6, 7]. The differences between men and women regarding the timing and extent of age-related changes are substantial [5, 6, 7]. All of these physical changes result in complex variations in the acoustic properties of speech (F0, stability of vocal fold vibration, spectral noise, speech tempo and formant frequencies), which are influenced by several variables (e.g., speaker-related factors or speech-material-related factors) [6].

A general finding is that older adults demonstrate a higher variation of acoustic features than younger adults. Most research has reported that F0 increases in males and decreases in females with aging [8]. Tremor and increased hoarseness, that have been associated with the aged voice, seems to be the result of a decline of F0 or amplitude stability [9]. It has also been noted that the age of the speaker strongly affects speaking and reading rates. Only relatively few studies have investigated age-related changes in the vocal tract resonance features. Most of them referred a general lowering of formants with age [10, 11] attributed to a lengthening of the vocal tract, caused by a lowering of the larynx, of the tracheobronchial tree and of the lungs, and by a growth of the facial skeleton. There also seems to be a trend towards vowel centralization [12, 8], but in some cases these changes have only been observed for particular vowels and speakers [12]. The longitudinal studies' findings (e.g., [13]) are in line with results of studies based on between-group comparisons (both F0 and F1 decrease with increasing age in the same speaker).

Additionally, changes in speech rhythm of old speakers have been scarcely examined [14].

For European Portuguese (EP) there is almost no data on the acoustic correlates of aging. Pellegrini et al. [14] conducted the first attempt to identify the most salient differences between older and younger adult speech in terms of acoustic features for EP, mostly to understand their impact on speech recognition performance. In a pilot study, our team [15] addressed effects of age and gender on formant frequencies of oral vowels produced by EP old speakers. We also investigated acoustic characteristics of EP children's speech [16]. So far the few studies on EP vowels acoustics [17, 18] have only provided information about vowels produced by young adult speakers.

Even though acoustic methods are a valuable tool in the study of speech sounds, the integration of acoustic with articulatory data could facilitate a comprehensive account of

10.21437/IberSPEECH.2021-29

anatomic–acoustic relationships. However, the existing studies mainly focused on the developmental period from infancy to adulthood [19]. Many instrumental techniques have been used in this field (e.g., Magnetic Resonance Imaging - MRI and ultrasound - US). US imaging is currently gaining popularity as a research tool in speech production studies because it is safe, non-invasive, simple and cost-effective. Moreover, it allows a simple synchronized acquisition of data (e.g., with audio). Although, acquisition and analysis of articulatory data is extremely demanding and poses several challenges.

The aim of the VoxSenes project is to study the speech changes due to aging, both at the segmental level as well as at suprasegmental level, by analyzing the variation of acoustic and articulatory parameters. A deeper knowledge on how speech changes with age is essential for the development of automatic speech recognition systems suitable for older voices (personalized reading aids and voice prostheses) and to clinical assessment of speech disorders.

2. Methodology

To achieve the defined aims, the VoxSenes project was organized in three different phases. In the first phase of this project, following up on pilot studies carried out by our team, the acoustic EP vowels produced by a large number of speakers, divided into different age groups, were analyzed. The relationship between the production of vowels (tongue shape and motion), the changes on formant frequencies and age were addressed in an ultrasound study to be held in the second task of the project. The articulatory basis of previously acoustic findings concerning speech aging were studied through the use of US tongue imaging synchronized with audio. The age-related changes of rhythm were approached in a third phase of the project, with the purpose of evaluating rhythmic changes throughout a certain span of time for a same individual (longitudinal study). The methodological procedures of phase 1 and 2 involve the recruitment of participants of different age groups. The procedures were submitted to Ethical Approval. Participants were fully informed about the goals, procedures and risks involved in research and had to give their consent to participate (Informed Consent). The researchers also guarantee the confidentiality and anonymity of the data, which means that identifying information will not be made available to anyone who is not directly involved in the study. The inclusion criteria were: age over 18 years old; European Portuguese as mother tongue. The exclusion criteria were: previous history of speech-language impairments, head/ neck cancer and/or neurological disorders; upper respiratory tract infection; currently smokers or had smoked within the previous 5 years; poor general health; and use of hearing aids. The participants were recruited through personal contacts and/or snowball technique in the community, and in Senior Universities from the center of Portugal.

2.1. VoxSenes Phase 1

The speech sample consisted of 36 real words, with the vowels of the EP [i], [e], [ɛ], [a], [o], [ɔ], [u] in stressed position and the vowels [ɐ] and [ɨ] in unstressed position. For each vowel, four different words were selected. Each vowel was produced in a disyllabic sequence, mostly CV.CV (C-consonant, V-vowel) where C was a voiced/ voiceless stop consonant or a voiced/ voiceless fricative consonant. The stimuli were embedded in a carrier sentence. The participants were also instructed to describe the "Cookie Theft" picture [20, 21] from the Boston Diagnostic Aphasia Examination at comfortable pitch and loudness level. All the recording took place in quiet rooms, using an AKG C535 EB cardioid condenser microphone connected to an external 16-bit sound system (PreSonus Audio-BoxTM USB), at a sampling rate of 44100 Hz. The sentences were randomized and presented individually using the software SpeechRecorder [22] with pictures and the orthographic word simultaneously. The recruited participants read the sentences 3 times at comfortable pitch and loudness level, after familiarizing with the sentences.

The recorded data from vowel production were automatically segmented at phoneme level using WebMAUS General for Portuguese language (PT) [23] and then imported into Praat [24]. Four trained analyzers manually checked the accuracy of the vowel boundaries. The acoustic parameters of the vowels - F0, median F0, F1, F2 and duration - were automatically extracted from the data set using Praat scripts. The recorded data from the picture description task were segmented for pauses over length 250 ms using a Praat scipt [25]. The Praat scripts ProsodyDescriptor [26] and the BeatExtractor [27] were used to extract the acoustic measures: total speech duration, percent pause time, mean pause duration, speaking rate, articulatory rate, speaking F0 and harmonic-to-noise ratio (HNR).

The research team is now determining the potential usefulness of the dynamic acoustic proprieties of EP vowels as carriers of age classification, a work that followed the one already carried out on the age effects in static cues (formant frequencies and F0) [15, 28, 29].

2.2. VoxSenes Phase 2

The synchronous acquisition of US images and speech sounds using Articulate Assistant Advanced software (AAA) [30] took place in a quiet room, using an endocavitary probe (65EC10EA) with 90º field of view positioned under the participants' chin using a stabilization helmet [31]. US was collected using a Mindray DP6900 at a frame rate of 60 Hz. Audio was collected with a Philips SBC ME400 microphone connected to an external sound system (UA-25 EX USB). The corpus consisted of 9 repetitions for each of the 9 EP oral vowels ([i], [e], [ɛ], [a], [o], [ɔ], [u], [ɐ] and [ɨ]). Corpus acquisition began with the production of the sequence /tatatata/ to assess sound and image synchronization and with swallowing saliva for hard palate delineation. The recorded data was collect as video and audio synchronized with SyncBrightUp unit [32], and the audio was automatically segmented, at phoneme level, using WebMAUS General [23].

Based on the acoustic midpoint of the vowels, the corresponding images were selected and processed using an unsupervised method to extract points-of-interest in the tongue. The method uses a radial sweep approach [33] and collects all the pixel intensities. The highest intensity point is extracted and the highest y coordinate is considered to represent the highest point of the tongue. The x-coordinate reflects the front-back position of the tongue in the x-y coordinate system. To allow intra- and inter-subject comparisons, normalization procedures were implemented. Concerning the analysis, the researchers intend to implement qualitative and quantitative analysis. The tongue surface tracing will be submitted to smoothing spline ANOVA (SSANOVA) [34] to find the best-fit curve across repetitions, allowing comparisons of tongue contours.

2.3. VoxSenes Phase 3

Longitudinal data of EP speakers (e.g., politicians or journalists) was obtained. The speech data were selected from national television archives. The selection of speech material for analysis depend mainly on the following criteria: similar type (e.g., reading or interviews) and duration of speech samples; quality of recordings and homogeneity of recording type across speakers and samples; voice quality of the speaker for enabling acoustic analysis.

For all sample, the vowel onsets were marked semi-automatically using the Praat script BearExtractor [27]. Fourteen rhythm and intonation parameters were extracted using the Praat script ProsodyDescriptor [26].

3. Results

3.1. VoxSenes Phase1

The results of Phase 1 were published, so far, in 1 journal paper ([35]) and 2 conference papers ([28]; [36]). The dynamic vowel studies result in one poster [29] and one position paper [37].

113 native Portuguese speakers (56 men and 57 women) from the central region of Portugal, aged between 35 and 97, participated in this phase of the VoxSenes project. The participants were divided into 4 age groups: [35-49] (15 men, 15 women), [50-64] (15 men, 15 women), [65-79] (15 men, 16 women), and ≥ 80 (11 men, 11 women).

The acoustic data revealed that the duration of all vowels increased with aging, and that the older speakers presented the longest vowel duration. F0 decreased in male until the age group [50-64] and started to increase after that age, with a more pronounced increase in the group ≥ 80, which presented the highest mean value of F0. For female speakers, the opposite tendency was observed, with an F0 increase until the age group [50-64] and a sharp decrease after this age. The age group ≥ 80 presented the highest mean value of F0. We also observed a general lowering of F1 and F2 frequencies for women in all stressed vowels; for males, changes observed in F1 and F2 were consistent with vowel centralization. There was no evidence of F3 decrease with age.

Regarding rhythm and intonation study, the most consistent age-related effects were an increase in speech pauses (both duration and percent time), mainly in men, and a HNR decrease in women. The articulatory rate differences between male and female decreased with age. Speaking F0 presented a similar tendency observed in the acoustic vowel data for male, with a decrease until the age group [65-79] and a rise after that age.

Concerning vowel dynamics, the results already achieved showed that dynamic measurements of F1-F3 result in better classification performance of senior/non-senior. Duration was also reconfirmed as an important predictor of age, just like in other studies of our team with static cues [35].

3.2. VoxSenes Phase2

The results already achieved on the articulatory changes in lifetime concern the accuracy of the method to extract tongue contours. The results were published in [38] and [39].

Data revealed that, in general, the method developed presented high accuracy mainly for the central regions of the tongue, where the highest point of the tongue is typically located. This method is currently being used, in an ongoing study with a larger sample, to characterize the articulatory changes of vowels with aging.

3.3. VoxSenes Phase3

Concerning phase 3 of the VoxSenes project, the researchers are currently selecting the speech data samples, more specifically interviews of two males in different ages, to allow the analysis of age effect in rhythm and intonation parameters.

4. Impact

Vocal aging is a multidimensional concept, making the study of older speech complex and challenging. The VoxSenes project intends to provide an opportunity for trying out a number of acoustic and articulatory research methods, in order to contribute to the increase of the phonetic knowledge concerning the properties of speech and its changes with age.

The results already achieved intend to have an impact on a better understanding of cross-linguistic similarities and in language-specific features of vowel aging. Furthermore, the new databases created in this project allow the characterization of vowel production in the normative aging process, being a reference for clinical assessment and intervention of different speech disorders. Concerning speech technology, data collected on the current project provides information that can have a positive impact on better age recognizers or classifiers, as well as in more natural-sounding synthesis of speaker age, which could be useful to improve the quality of life of older people [40, 41].

5. Acknowledgements

This research project is financially supported by POCI-01-0145-FEDER-03082 (funded by FEDER, through COMPETE2020 - Programa Operacional Competitividade e Internacionalização (POCI), and by national funds (OE), through FCT/MCTES). The researchers also intend to acknowledge the grant SFRH/BD/115381/2016 and IEETA (UIDB/00127/2020). We are very grateful to the institutions for having made possible the data collection, and also to all adults who contributed as speakers.

6. References

[1] W. He, D. Goodkind, and P. R. Kowal, "An aging world: 2015," *International Population Reports*, vol. P95/16-1, 2016.

[2] Statistics Portugal, "Estimativas de População Residente em Portugal - 2018 (Estimates of resident population in Portugal - 2018)," *Destaque: informação à comunicação social*, 2019.

[3] ——, "Envelhecimento da população residente em Portugal e na União Europeia (Aging of the resident population in Portugal and the European Union)," *Destaque: informação à comunicação social*, 2015.

[4] P. Massimo and P. Elisa, "Age and Rhtymic Variations: A study on Italian," in *INTERSPEECH 2014*. Singapore: ISCA, 2014, pp. 1234–1237.

[5] S. E. Linville, *Vocal aging*. Australia, San Diego: Singular Thomson Learning, 2001.

[6] S. Schötz, *Perception, analysis and synthesis of speaker age*. Lund University: Linguistics and Phonetics, 2006, vol. 47.

[7] K. Makiyama and S. Hirano, "Aging Voice," Singapore, 2017.

[8] P. Torre III and J. A. Barlow, "Age-related changes in acoustic characteristics of adult speech," *Journal of Communication Disorders*, vol. 42, pp. 324–333, 2009.

[9] S. E. Linville, "The Sound of Senescence," *Journal of Voice*, vol. 10, no. 2, pp. 190–200, 1996.

[10] P. J. Watson and B. Munson, "A comparison of vowel acoustics between older and younger adults," in *ICPhS XVI*, Saarbrücken, 2007, pp. 561–564.

[11] S. A. Xue and G. J. Hao, "Changes in the Human vocal tact due to aging and the acoustic correlates of speech production: a pilot study," *J Speech Lang Hear Res*, vol. 46, no. 3, pp. 689–701, 2003.

[12] M. P. Rastatter, R. A. McGuire, J. Kalinowski, and A. Stuart, "Formant frequency characteristics of elderly speakers in contextual speech," *Folia Phoniatrica et Logopaedica*, vol. 49, no. 1, pp. 1–8, 1997.

[13] J. Harrington, S. Palethorpe, and C. I. Watson, "Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers," in *INTERSPEECH*, Belgium, 2007, pp. 2753–2756.

[14] T. Pellegrini, A. Hämäläinen, P. B. de Mareüil, M. Tjalve, I. Trancoso, S. Candeias, M. S. Dias, and D. Braga, "A corpus-based study of elderly and young speakers of European Portuguese: acoustic correlates and their impact on speech recognition performance," in *INTERSPEECH*, Lyon, 2013, pp. 852–856.

[15] L. Albuquerque, C. Oliveira, A. Teixeira, P. Sa-Couto, J. Freitas, and M. S. M. Dias, "Impact of age in the production of European Portuguese vowels," in *INTERSPEECH*, Singapore, 2014, pp. 940–944.

[16] C. Oliveira, M. M. Cunha, S. Silva, A. Teixeira, and P. Sa-Couto, "Acoustic analysis of European Portuguese oral vowels produced by children," in *IberSPEECH*, vol. 328, Madrid, Spain, 2012, pp. 129–138.

[17] M. R. D. Martins, "Análise acústica das vogais orais tónicas em Português," *Boletim de Filologia*, vol. 22, pp. 303–314, 1973.

[18] P. Escudero, P. Boersma, A. S. Rauber, and R. A. H. Bion, "A cross-dialect acoustic description of vowels: Brazilian and European Portuguese," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1379–1393, 2009.

[19] H. K. Vorperian and R. D. Kent, "Vowel Acoustic Space Development in Children: A Synthesis of Acoustic and Anatomic Data," *J Speech Lang Hear Res*, vol. 50, no. 6, pp. 1510–1545, 2007. [Online]. Available: http://jslhr.asha.org/cgi/content/abstract/50/6/1510

[20] H. Goodglass and E. Kaplan, *The Assessment of Aphasia and Related Disorders*, 2nd ed. Philadelphia, PA.: Lea and Febiger, 1983.

[21] E. E. Morgan and M. Rastatter, "Variability of voice fundamental frequency in elderly female speakers." *Perceptual and motor skills*, vol. 63, no. 1, pp. 215–218, 1986.

[22] C. Draxler and K. Jänsch, "SpeechRecorder (3.12.0)," 2017.

[23] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech and Language*, vol. 45, pp. 326–347, 2017.

[24] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," University of Amsterdam, 2012. [Online]. Available: http://www.praat.org/

[25] N. H. de Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior Research Methods*, vol. 41, no. 2, pp. 385–390, may 2009.

[26] P. A. Barbosa, "Semi-automatic and automatic tools for generating prosodic descriptors for prosody research," in *TRASP*, vol. 13, no. 2, Aix-en-Provence, 2013, pp. 86–89.

[27] ——, "Automatic duration-related salience detection in Brazilian Portuguese read and spontaneous speech," in *Speech Prosody*, Chicago, 2010.

[28] L. Albuquerque, C. Oliveira, A. Teixeira, P. Sa-Couto, and D. Figueiredo, "Age-related changes in European Portuguese vowel acoustics," in *INTERSPEECH*, Graz, Austria, 2019, pp. 3965–3969.

[29] L. Albuquerque, A. Teixeira, C. Oliveira, and D. Figueiredo, "The effect of dynamic acoustic cues on age classification," in *SPPL2020: 2nd Workshop on Speech Perception and Production across the Lifespan (Poster)*, 2020, p. 81.

[30] Articulate Assistant Ltd., "Articulate Assistant Advanced ultrasound module user manual," 2014.

[31] Articulate Instruments Ltd., "Ultrasound stabilisation headset users manual," Edinburgh, UK, 2008.

[32] ——, "SyncBrightUp users manual," Edinburgh, UK, 2010.

[33] L. Ménard, C. Toupin, S. R. Baum, S. Drouin, J. Aubin, and M. Tiede, "Acoustic and articulatory analysis of French vowels produced by congenitally blind adults and sighted adults," *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 2975–2987, 2013.

[34] L. Davidson, "Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 407–415, 2006.

[35] L. Albuquerque, C. Oliveira, A. Teixeira, P. Sa-Couto, and D. Figueiredo, "A comprehensive analysis of age and gender effects in European Portuguese oral vowels," *Journal of Voice*, no. In press, dec 2020.

[36] L. Albuquerque, A. R. S. Valente, A. Teixeira, C. Oliveira, and D. Figueiredo, "Acoustic changes in spontaneous speech with age," in *VIII Congreso Internacional de Fonética Experimental*, Girona, 2021.

[37] L. Albuquerque, C. Oliveira, A. Teixeira, and D. Figueiredo, "Eppur si muove: Formant dynamics is relevant for the study of Speech Aging Effects," in *14th BIOSIGNALS*, Online, 2021, p. in press.

[38] F. Barros, A. R. Valente, L. Albuquerque, S. Silva, A. Teixeira, and C. Oliveira, "Contributions to a quantitative unsupervised processing and analysis of tongue in ultrasound images," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12132 LNCS, pp. 170–181, 2020.

[39] F. Barros, S. Silva, L. Albuquerque, A. R. Valente, A. Teixeira, P. Martins, and C. Oliveira, "Towards the use of ultrasonography to study aging effects in vowel production," in *12th ISSP*, Online, 2020, p. Poster.

[40] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "Speaker age estimation on conversational telephone speech using senone posterior based i-vectors," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5040–5044. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7472637/

[41] M. Yue, L. Chen, J. Zhang, and H. Liu, "Speaker age recognition based on isolated words by using SVM," in *CCIS2014*, 2014, pp. 282–286.

Hispabot-Covid19: the official Spanish conversational system about Covid-19

David Griol[1], *Zoraida Callejas*[1], *David Pérez-Fernández*[2]

[1]Department of Languages and Computer Systems, University of Granada,
Periodista Daniel Saucedo Aranda s.n., 18071, Granada, Spain
[2]Ministry of Inclusion, Social Security and Migration,
Paseo de la Castellana 63, 28071, Madrid, Spain
dgriol@ugr.es, zoraida@ugr.es, david.perez@inv.uam.es

Abstract

Hispabot-Covid19 is a conversational system developed for the Spanish Government to provide responses to frequently asked questions related to the pandemic originated by Covid-19 and its implications in Spain. The system received more than 350,000 queries between April and June 2020, being a clear example of how conversational systems can be applied to reduce the pressure on the health emergency phone lines and to provide 24/7 access to information and services using natural language. In this paper, we describe the main features of the Hispabot-Covid19 system.
Index Terms: Covid-19, Conversational Systems, FAQ, Spanish.

1. Introduction

The Secretary of State for Digitalization and Artificial Intelligence (SEDIA)[1], through the Spanish Plan for the Advancement of Language Technology[2], coordinated in March 2020 the development of a conversational assistant to answer frequent questions about COVID-19.

The Hispabot-Covid19 conversational system [1] accessed information from official sources from the Spanish Ministry of Health and the World Health Organization to report on various issues, such as: symptoms, vulnerable groups, transmission, prevention, coexistence with infected people, conditions for quarantine and isolation, attention telephone numbers, among many others. The assistant also incorporated the information published in the Official Spanish State Bulletin (Boletín Oficial del Estado, BOE) regarding the application of the *State of Alarm* and the *Plan for the Transition towards a New Normality*. The assistant did not retrieve or analyze any personal data.

Hispabot-Covid19 is an example of the application of conversational systems as a solution to provide information to citizens and facilitate contact with public institutions [2, 3, 4]. This objective can be achieved with:

1. systems for solving frequent doubts related to public services, so that it is possible to have automated FAQs where questions can be asked in natural language. In this case the questions are isolated, it is not necessary to take into account the questions previously asked by the citizen.
2. systems that can hold a dialogue with the citizen to automate services, e.g. to make arrangements with an administration. In this case it is necessary to have systems that process the previous history of the dialogue.

[1]https://avancedigital.mineco.gob.es (Last access: February 2021)
[2]https://plantl.mineco.gob.es (Last access: February 2021)

Hispabot-Covid19 falls within the first group. For the development of the system it has been necessary to develop the different modules and models, deploy it in several interaction channels (web-based assistant, Telegram and WhatsApp), and optimize the system on the basis of the observed interactions. For the development phase, a model was built for language processing. For this purpose, a list of the main entities or concepts handled (e.g. autonomous communities, synonyms of coronavirus, names of related diseases, etc.) has been elaborated and, on the other hand, a training of the natural language understanding model with an initial corpus of training phrases was carried out.

All dialogues have been recorded in the system and processed for continuous retraining and improvement of the assistant. To this end, new training phrases and new question categories were incorporated daily, misunderstandings have been detected and corrected, and the knowledge base has been restructured according to updates in the information provided.

On April 3rd 2020, the Ministry of Economic Affairs and Digital Transformation launched a pilot project to firstly integrate the assistant into the web portals of the Government of La Rioja. From April 3rd to 7th, the assistant received more than 5,000 queries on these portals. From April 8th until the end of June 2020 it was accessible through WhatsApp and Telegram. More than 350,000 interactions took place during this period.

In this paper we describe the main features of the Hispabot-Covid19 system, the main phases that were followed to develop the system, the set categories of FAQs that were defined and the main statistics related to the operation of the system.

2. State of the art

Establishing a natural, agile and fluent conversation with a machine using natural language has been one of the main research challenges in the fields of Natural Language Processing, Computational Linguistics and Speech Technologies. This challenge has captured since years ago the interest in the academic, commercial and industrial fields, especially considering the wide range of applications of this kind of systems [2].

In fact, several recent reports highlight how the use of natural language (and mainly voice) is changing the way we relate to technology, with growth prospects in technologies, sectors and devices that indicate that we are in the ”era of conversational interfaces” and that have been accentuated during the pandemic caused by Covid-19 with the increased use of e-government.

The reference legislation on e-Government in Spain dates back to June 2007. The circumstances of the pandemic have prevented the staff of the Public Administrations from performing their functions in the offices, have generalized the implementation of teleworking tools and have also made possible a

10.21437/IberSPEECH.2021-30

Figure 1: *Screenshots of the integration of the Hispabot-Covid19 in the website of La Rioja Salud*

generalized use of telematic means by the citizens. Information and communication technologies have played a fundamental role in maintaining the activity of the Public Administration during this crisis, and have shown the future importance of continuing with the digital transformation.

The use of electronic media has also proved to be a magnificent opportunity to accelerate the number of developments and the application of conversational systems in different areas of the Administration, both to facilitate access to information and answers to frequently asked questions, as well as to carry out e-government procedures.

Chatbots have become a tool to combat misinformation and even to channel people's anxiety about possible contagions. The use of chatbots as a strategy to deal with the pandemic has been both nationally and globally. Regarding the resolution of frequently asked questions, it is worth noting the large number of conversational systems that have been developed in a very short time worldwide to answer questions related to the pandemic caused by Covid-19, some of them also including triage services related to the symptoms of the disease: Hispabot-Covid19 [1], Carina[3], COVID19AragónBot[4], IMPAI[5], World Health Organization chatbot[6], the UCSF Health and Northwell Health chatbot[7], the CDC chatbot[8], CovidBot[9], Boti[10], etc. The usage statistics of these assistants have shown that they have become indispensable global tools in the fight against the virus, in combination with the Web channel and messaging services as an ideal environment to carry out action protocols, provide information and raise awareness about the pandemic generated by COVID-19.

3. The Hispabot-Covid19 system

As described in the introduction section, since the beginning of the pandemic, official State agencies have made a great effort to provide data and information in real time and to reduce the collapse of the citizen health telephone lines. The Hispabot-Covid19 system has contributed to this effort by providing easy access to official information, allowing citizens to ask questions in their own words, and providing a practical example of the great potential that these systems will have for communication between citizens and the Public Administration.

For the development of the system it has been necessary to work on 4 main lines:

- Supervision of contents: The Health Service of La Rioja

[3]https://1millionbot.com/chatbot-coronavirus/ (Last access: February 2021)

[4]https://www.europapress.es/aragon/noticia-itainnova-pone-marcha-covid19aragonbot-telegram-responder-dudas-ciudadanos-20200403100820.html (Last access: February 2021)

[5]https://campussanofi.es/e-professionals/noticias/impai-chatbot-coronavirus/ (Last access: February 2021)

[6]https://www.whatsapp.com/coronavirus/who (Last access: February 2021)

[7]https://www.healthcareitnews.com/news/northwell-ucsf-unc-using-chatbot-and-related-tech-manage-covid-19-patients (Last access: February 2021)

[8]https://www.theverge.com/2020/3/21/21189227/cdc-microsoft-chatbot-coronavirus-symptom-checker (Last access: February 2021)

[9]https://www.infobae.com/tecno/2020/04/01/asi-es-el-chatbot-mexicano-que-lucha-contra-la-desinformacion-sobre-coronavirus/ (Last access: February 2021)

[10]https://planetachatbot.com/buenos-aires-lanza-boti-chatbot-en-app-mas-utilizada-mundo-9ca18f2d5678 (Last access: February 2021)

Table 1: *Distribution of users' queries for each one of the intents defined for the Hispabot-Covid19 System*

Intent	Perc.	Intent	Perc.	Intent	Perc.
Visits to dental centers	0.12	Meetings during the lockdown	0.31	How to care for patients	0.48
Attendance to health centers	1.00	Rites and ceremonies	0.77	School - educational institutions	1.27
Effects in different ages	0.17	Questions without answer	2.38	Welcome and greetings	1.10
Droplet transmission	0.32	Patients with diabetes	0.23	Differences other diseases	0.28
Home isolation to avoid spread	0.78	Apologies for misunderstandings	0.69	Interpersonal distance	0.20
Nutritional advice	0.50	Official and technical documents	0.19	Donations	0.08
Hospital discharge	0.08	Blood donations	0.24	Duration of isolation for patients	0.20
Pets and transmission in animals	1.13	How to entertain children	0.31	Role of state security forces	0.12
Use of antibiotics	0.13	Pregnancy-related questions	0.50	Use of nasal swabs	0.05
Anticoagulated patients	0.12	Patients with epilepsy	0.11	Chronic respiratory diseases	0.09
Flattening the COVID-19 peak	0.25	Questions about the alarm state	0.69	Stigmatizing behaviors	0.11
Applause for healthcare workers	0.20	Explain the disease to children	0.37	Questions about pharmacies	0.23
Asymptomatic patients	0.17	Different phases and provinces	1.53	End date of lockdown.	3.12
Asthmatic and allergic patients	0.42	Time slots and allowed activities	3.43	Fruits and vegetables	0.25
Patients home care	0.16	Official sources of the contents	0.82	Thank you and goodbye	0.85
Self-employed workers	0.24	Questions on vulnerable groups	1.00	Recommendations for shopping	1.35
Emergency phone numbers	2.50	Transmission through feces	0.09	Patients with hypertension	0.42
When to use emergency phones	0.39	Tourism and accommodation	0.09	Questions about ibuprofen	0.23
Workers sick leave	0.62	How to use the system	0.36	Insults and bad language	0.23
Widespread fakes about Covid	0.80	Congresses and scientific events	0.13	Encouraging queries	0.57
Specific fake about using oil	0.04	Breastfeeding	0.20	Advice on hand washing	0.58
Specific fake about eating garlic	0.04	Washing face masks	0.23	Use of contact lenses	0.13
Specific fake about using alcohol	0.08	Use of the car and urban transport	1.55	Clean the shopping basket	0.32
Killing the virus with a hair dryer	0.04	Hygiene recommendations	0.86	Disinfection commercial spaces	0.08
Virus transmission through shoes	0.20	Disinfection of clothes	0.25	Days of national mourning	0.05
Specific fake of using UV light	0.07	Gender violence	0.31	Information about face masks	2.14
How to stop spreading fakes	0.27	Information purchase of masks	0.54	Medicines and vaccines	1.09
Influence of temperature, zones	0.28	Information on social measures	0.77	Information on virus mortality	0.09
Oncology patients	0.20	Using/cleaning mobile phones	0.33	Fines and associated regulations	0.37
COVID-19 and viral load	0.05	Not touching nose-eyes-mouth	0.05	Relation with pneumonia	0.13
Questions about cybersecurity	0.31	Access to official press releases	0.25	News and updates	0.48
Reported cases and deaths	5.51	Updates related to the system	0.11	Works, reforms and construction	0.20
How report. cases are calculated	0.36	Origin of the virus	0.29	Information on specific jobs	0.17
How the disease is spread	0.53	Meaning of pandemic	0.12	Use of paracetamol	0.31
Infection through water	0.15	Information about home orders	0.48	Specific information hairdressers	1.45
Transmission in the workplace	0.76	Incubation period of the virus	0.74	Information staying at home	0.23
Transmission by direct contact	0.36	Inform. recoverable paid leave	0.37	System's privacy policy	0.07
Neighborhood coexistence	0.64	Immunity and recovery	0.08	Inform. how to handle worry	0.28
Quarantine related questions	0.54	Private health care information	0.12	Inform. probability of infection	0.23
Inform. healthcare professionals	0.21	Protection and prevention meas.	1.49	Information about covid-19 tests	1.54
General inform. about disease	0.88	Inform. system's functionalities	1.29	What not to do	0.13
Providing the name of the bot	0.61	How to safely remove gloves	0.20	Antibodies and immunity	0.82
People with kidney conditions	0.12	Residences - sheltered housing	0.11	Agricultural and livestock activit.	0.77
People with autism spectrum	0.23	Bars, terraces and restaurant	0.76	Care of orchards and gardens	0.17
Recommend. for leaving home	0.48	Regulations for leaving home	2.68	Permissible duration of outings	0.21
Recommend. on children walks	2.97	Libraries, cinemas, theaters	0.35	Visits to beaches - green spaces	0.53
Making and receiving visits	1.22	Walking the dog	0.49	Garbage disposal	0.11
Work permits and safeguards	1.39	Use of public transport	0.48	Inform. related to mental health	1.41
General information about SARS	0.16	Seroprevalence	0.16	Share Hispabot on WhatsApp	0.32
Symptoms in children	0.65	How to act when symptoms	5.32	Essential supplies lockdown	0.21
Transmission of virus on surfaces	0.44	Information about smoking	0.19	Information about teleworking	0.31
Information about open stores	2.31	Information about scams	0.12	Inform. of working conditions	0.96
Administrative procedures	0.86	Interprovincial travel regulations	2.16	Treatment of the disease	0.27
Information on the use of gloves	0.35	Vaccine information	0.07	Incoming international travels	0.25
Outgoing international travels	1.06	Transmission area	0.19		

and the Ministry of Health participated in the revision and accessibility of the main contents provided by the conversational assistant.

- Development of the conversational system: A natural language understanding model was built using Google technologies. For this purpose, a list of 90 entities with

more of 2100 different as possible values for these entities were defined (e.g. autonomous communities, synonyms of coronavirus, names of related diseases, etc.). More than 8,000 training sentences were gradually incorporated to process a total of 164 intents (categories of questions that can be answered by the system). Table 1 shows the list of intents defined for the system. The numbers indicated in the table show the percentage of users' sentences that were classified in each category during the operation of the system. In Hispabot-Covid19 there was only one defined answer for each intent. These responses have been designed either as direct answers in text mode (including text, emojis and bold and italic highlighting) or defined after processing them using a script (e.g. to indicate the telephone number corresponding to the autonomous community indicated by the user). The most frequently asked questions corresponded with the hot topics of each day or the most prominent news in the media. For example, during the first weekend of April, the number of questions on the use of masks increased considerably, and when there was news of the infection of pets, the number of queries on the transmission of COVID-19 between humans and animals increased. The daily analysis of the queries provides very valuable information to know in real time which were the most frequent doubts of the citizens.

- Deployment in interaction channels: The system was deployed for the interaction in messaging services (WhatsApp and Telegram, Figure 1) and web environments (Figure 2) More than 350,000 interactions were recorded with the system from April to June 2020.
- System maintenance: The system was daily maintained to detect understanding failures, restructure the knowledge base according to the updates and incorporate entities, categories, and training sentences. Rates higher than 94% were achieved with regard the correct classification of users' queries in the correct intent. A set of scripts were developed to dump the data (intents and entities) to CSV format in order to have a readable version of the data for monitoring.

Figure 2: *Screenshots of the integration of the Hispabot-Covid19 assistant in WhatsApp*

4. Conclusions and future work

In an emergency situation such as the one we are currently experiencing, providing information, updates and support to citizens are key objectives. Covid-19 has accelerated the digitization process and made it urgent to provide timely and permanent support to citizens rooted in trustworthy sources of information.

In this context, the implementation of chatbots has experienced a significant increase and revealed the full potential of these systems, especially in the healthcare sector. They are a valid complementary help to provide information and services and lighten the workload of official phone numbers in emergency situations.

The Hispabot-Covid19 system has contributed to this aim by providing an easy access to official information, allowing citizens to formulate questions in their own words and providing a practical example of the great potential that these systems will have for communication between citizens and the Public Administration. The number of users' queries received by the system show the potential for the application of conversational systems in information and citizen services as consolidated in the Administration as the 012 telephone number (it received an average of 900,000 queries per year in the Community of Madrid between 2004 and 2007), the average of nearly 900,000 calls received monthly by the 060 service at present, or the recent application of these systems in the Social Security portal to provide a guide of steps to follow for the resolution of procedures and answer frequently asked questions.

The corpus acquired by means of the interaction of citizens with the Hispabot-Covid19 conversational system will be published in a open access repository in the next months with detailed instructions about its labeling (sentences IDs, entities, date and hour). We want to extend this information with the labeling of predicate-argument structures, the definition of different kinds of word embedding for the task, and the additional definition of train and test partitions.

5. Acknowledgements

We want to thank the Spanish Plan for the Advancement of Language Technology to promote the scientific and innovation potential of Spain in the field of Natural Language Technologies.

6. References

[1] D. Pérez, D. Griol, and Z. Callejas, "El sistema Hispabot-Covid19: sistemas conversacionales para la Administración y la Atención al Ciudadano," *Boletic*, vol. 86, 2020. [Online]. Available: Available in: https://www.astic.es/la-asociacion/boletic/boletic-no-86-julio-2020f

[2] M. McTear, Z. Callejas, and D. Griol, *The Conversational Interface: Talking to Smart Devices*. Springer, 2016.

[3] M. McTear, *Conversational AI. Dialogue systems, Conversational Agents, and Chatbots*. Morgan and Claypool Publishers, 2020.

[4] J. Quesada, Z. Callejas, and D. Griol, *Informe sobre sistemas conversacionales multimodales y multilingües*. Plan de Impulso de las Tecnologías del Lenguaje. Available in: https://plantl.mineco.gob.es/tecnologias-lenguaje/actividades/estudios/Paginas/sistemas-conversacionales.aspx, 2019.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Project MEMNON: Extending Speech Production Studies to Silent Speech, Dynamic Sounds and Audiovisual Speech Synthesis

Samuel Silva[1], *António Teixeira*[1], *Nuno Almeida*[1], *Diogo Silva*[1], *David Ferreira*[1], *Conceição Cunha*[2]

[1]IEETA, DETI, University of Aveiro, Aveiro, Portugal
[2]Institute of Phonetics and Speech Processing, LMU Munich, Germany
sss@ua.pt, ajst@ua.pt

Abstract

This paper presents ongoing research project MEMNON. To move beyond the current state-of-the-art for speech production (SP) studies, and profiting from the team's strong background in this field, this project advances the body of knowledge regarding SP of European Portuguese, based on multimodal (e.g., rtMRI, US) acquisition, processing, and analysis of speech production data, and applies novel and existing knowledge to propose innovative contributions to articulatory-based audiovisual speech synthesis and silent speech interfaces, serving both research and interaction.

Index Terms: speech production studies real-time MRI, data-driven analysis, audiovisual speech synthesis, silent speech interfaces

1. Context and Motivation

The use of speech, our natural form of communication, is expected to improve our interaction with machines, but, its efficient use for interaction, challenges state-of-the-art speech technology and the available knowledge on Human-Human and Human-Machine communication.

Nowadays, information regarding SP can be obtained using a wide range of technologies such as electromagnetic midsagittal articulography (EMA), magnetic resonance imaging (MRI), real-time MRI (rtMRI), ultrasound (US), and surface electromyography (EMG). These technologies provide resources to advance many subareas of speech research, from Phonetics to speech synthesis/recognition and silent speech interface (SSI), benefiting, e.g., the speech impaired, and contributing to the definition of normality in health applications (e.g. speech therapy), improving diagnostic/ intervention. Among the different technologies supporting speech production studies (SPS), rtMRI [1] and US [1] are increasingly used and result in large amounts of data to be processed. For imaging techniques, analysis mostly relies on subjective evaluations (e.g. visual comparison of vocal tract contours) hindering replicability and computational approaches. Without a proper approach to these aspects, SPS tend to provide results below the potential of the technologies used. Therefore, in the context of SPS, it is very important to propose methods to handle the data, providing the extraction and quantitative assessment of relevant features and addressing challenges concerning the access to high dimensional datasets, storage/retrieval of the collected and generated data and the proposal of adequate processing and analysis methods. Tackling these challenges is a precondition for advancing the stateof-the-art in SPS.

2. Challenges

Considering the state-of-the-art and our previous contributions to the field [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17], it is possible to identify a set of problems and open questions that were considered at the onset of this project:

- Although the articulatory characterization of EP oral and nasal vowels has been addressed, the characterization of EP sounds (e.g., vocoids) is lacking for complex sounds such as EP diphthongs;
- Audiovisual speech synthesis (AVS) can be an important asset for speech research and interaction, but current trends on data-driven approaches do not serve an SP research setting. Also, visual coarticulation still requires attention;
- SSI based on non-invasive technologies, such as video, is a promising approach, but use of additional knowledge regarding SP, such as articulatory events, or other modalities (e.g., ultrasonic Doppler) might improve performance.

The problems identified above pose a set of challenges. The first concerns which technologies, and how, should be considered to expand our knowledge on EP sounds. For instance, regarding EP diphthongs, while rtMRI is more expensive than US, the latter only allows studying oral configurations. So, it is important to devise methods that combine data from different modalities in such a way that all relevant (and available) aspects are covered.

A second challenge concerns how to deal with all the data required to address the problems identified. To advance the state-of-the-art and address new problems, processing and analysis methods targeting articulatory data need to further support quantitative systematic multimodal analysis. This should also support researchers who may lack the processing skills, but could play a key role in exploring and gathering additional insight over the data if given the proper tools, e.g. linguists.

Regarding AVS, data-driven approaches perform well, but seldom contribute to advance our knowledge on SP. The challenge is how to address the complexity of a common model for both speech and visual output and avoid ad hoc treatment of coarticulation.

Additional challenges arise from the ground-breaking nature of the required research. For SSIs, the only works on the literature addressing EP are those by Freitas et al., e.g. [12, 13, 15, 18], and much work is yet to be done regarding the exploration of non-invasive technologies.

3. Project MEMNON

To tackle the challenges identified above, our team proposed project MEMNON. Its main goals and overall approach are summarized in what follows.

10.21437/IberSPEECH.2021-31

3.1. Main Areas of Actuation

Overall, project MEMNON is grounded on the pursuit of research that advances our knowledge on the more basic aspects, contributing to speech science, but motivated and challenged by its application in real scenarios. Therefore, this project reflects such vision by considering contributions in three different levels: 1) consideration of multiple technologies to obtain speech data and advance the body of knowledge for EP sounds description; 2) consideration of articulatory descriptions to advance science for SP modeling, oriented to synthesis, working to support research; and 3) taking articulatory data and modeling a step forward to propose advanced speech-based human computer interaction technologies relevant for different scenarios such as Speech Therapy and Ambient Assisted Living.

3.2. Overall Structure and Methods

MEMNON relies on an interdisciplinary approach combining researchers, knowledge and methods from Speech Science, Interaction, Image Processing, Phonetics, Imagiology, Mathematics, Visualization and Computer Science. The research is organized in five parallel activities:

- **Data Acquisition** — Our approach to the problems and challenges described above is grounded on the acquisition of novel databases covering different aspects of SP in EP. This workpackage defines the corpora to consider and technologies (e.g., rtMRI, US), based on initial requirements set by the remaining lines of research.
- **Data Processing and Analysis Methods** — explores novel methods to process the acquired data (rtMRI, US, video, etc.) as required by the remaining activities (tongue positioning, velar dynamics, etc). Besides investing on novel methods to extract features from the data (e.g., tongue contour in US images), one important goal is to propose and test statistical and data-driven methods [] for the automatic determination of important articulatory features such as critical articulators and gestures.
- **Study of Dynamic Sounds** — This will profit from the new SP databases and the main goal is to serve the study of EP oral and nasal diphthongs. Vowels, particularly oral ones, will also be included to allow comparisons. As some of the diphthongs are unstable, dependent on the speed of articulation, at least two different rates - one faster than the usual production of the subject - will be contemplated.
- **Audiovisual Speech Synthesis** — We take our experience in articulatory speech synthesis [], capitalize on the gestural descriptions provided by the research on dynamic sounds, and propose that audiovisual speech synthesis can be performed by using the articulatory parameters (considered for articulatory speech synthesis) to animate an advanced 3D avatar obtained from the company Face in Motion during a technological transfer in the scope of Marie Curie IAPP project IRIS (ref. 610986) [19] and currently in the market by the company MyDidimo. Furthermore, we propose, system should not be proposed isolated, but integrated in a framework including an interaction modality, to ease its deployment, and an evaluation module supporting the implementation of different methods (e.g. perceptual tests and unsupervised evaluation) to perform validation and inform development.
- **Silent Speech Interfaces** — will acquire a new multimodal dataset contemplating non-invasive technologies e.g., RGBD video. This new data, based on corpora specifically tailored to cover different aspects of EP, will allow assessing the complementarity of the different technologies. Based on technologies such as US, used as ground truth for the movement of the tongue, non-invasive technologies can be explored for their ability to provide articulatory data and fused to improve recognition performance. Finally, a prototype articulatory-SSI system will be developed and tested.

4. Results Overview

This section presents an overview of the results obtained, so far, in the scope of project MEMNON. For a more detailed description of methods and outcomes, the reader is forwarded to the authors' references provided along the text.

4.1. New speech production data

To go beyond our previous research and, particularly to address the study of the dynamics and coordination of nasals and diphthongs, data from the whole vocal tract is needed at a higher frame rate than our previous data (14fps).

In a collaboration with the Institute of Phonetics and Speech Processing, LMU Munich, Germany, under the scope of project 'Synchronic variability and change in European Portuguese', led by Conceição Cunha, we have been able to work on the definition and acquisitions of a novel high frame rate RT-MRI dataset.

The recordings were performed at the Max-Planck-Institute in Göttingen, Germany, using a 3T Siemens Prisma Fit MRI System equipped with a 64-channel head coil. The MRI acquisitions involved a low-flip angle gradient-echo sequence with radial encodings and a high degree of data under sampling [20]. The procedure allowed for real-time image sequences of the vocal tract in a midsagittal plane of the speaker at 50 fps. Speech sound was synchronously recorded using an optical microphone.

The corpus consists of lexical words containing: 1) all EP oral and nasal vowels in one/two syllable words; 2) oral and nasal diphthongs including alternations of nasal monophthongs and diphthongs as in *'som'* ('sound') and *'são'* ('they are'), recorded for further investigation of variability in the production of nasality.

4.2. MRI processing platform

MEMNON's MRI processing platform [21] aims to support speech production studies by providing researchers support to: 1) organize, annotate and process data; 2) off-the-shelf processing components, this way allowing them to create more complex processing pipelines; 3) collaboration and distribution of tasks (e.g., data revision) among multiple speech researchers; and 4) allow a ubiquitous access to the resources.

The platform's architecture is presented, to the left, in Figure 1. The platform runs in the cloud, allowing users to access it almost from any computational device.The user interacts with the platform throw a back-end module, this module manages interaction between user and system. It enables the access to the processing pipeline and data analyses. The Processing pipeline, executes all the intensive computational operation, such as image segmentations. The data analyses executes the tasks need to extract meaningful information from segmented data, e.g. distances between articulators. At this point, and profiting from a large set of already annotated and revised images, other solutions for image segmentation are being considered exploring machine learning approaches.

In its current state, the platform allows speech production researchers to upload acquisitions of RT-MRI data and organize

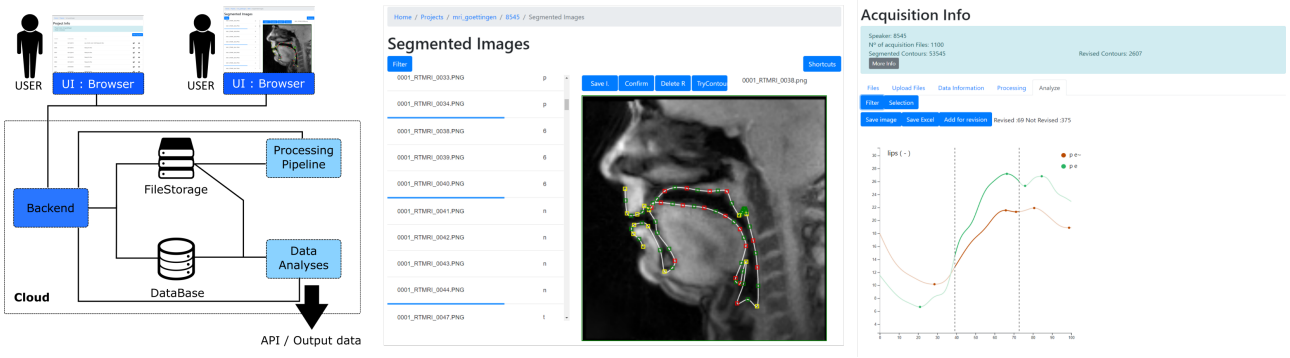


Figure 1: *Details regarding MEMNON's MRI processing and analysis platform. From left to right: overall architecture, visualization and revision of vocal tract profile, and analysis of articulator movement, over time.*

it into projects. It also provides methods for the semi-automatic extraction of the vocal tract outline by requiring the manual segmentation of a small set of MRI images which the platform then uses to train a model and segment the complete set of images [22]. In this regard, the platform also provides users with a simple method to revise the automatic segmentation (see Figure 1, at the center), supporting the attribution of revision tasks to different users of the platform.

At this stage, the platform also provides first versions of methods for data analyses, e.g., lip or velopharyngeal passage aperture along the production of a particular sound (see Figure 1, on the right). One important aspect regarding this latter feature is the possibility to filter the considered data based on the desired targets, contexts, and words of the corpus.

Finally, and to enable a more versatile use of the data, e.g., for prototypng of novel methods using other tools, such as Matlab or R, the platform also provides services to retrieve raw data based on specific filters. An API responds to requests done by third party scripts, with a filtered list of the requested segmentation in JSON format.

4.3. Analysis of Speech Production Data

The resulting speech production data creates a set of challenges concerning how to deal with the its complexity and size, but also a range of opportunities to explore data-driven methods for speech production studies. Our first approaches to obtain the vocal tract data from the RT-MRI images are grounded on earlier work regarding vocal tract segmentation from RT-MRI [22]. The following sections provide a short overview of MEMNON's contributions regarding data-driven analysis of speech production data.

4.3.1. Critical Articulators

The study of critical articulators (and gestures) is relevant for different aspects of our research. First, it informs advancing our previous work regarding the phonological descriptions of EP sounds [23] and, second, it provides information that can be used to model EP sounds in articulatory synthesis [4].

In earlier work [24], the authors demonstrated that critical articulator could be identified by extending the applicability of the method proposed for EMA data by Jackson and Singampalli [25] to MRI data. Those first results were obtained for RT-MRI at 14Hz and, since the original method worked with 100 Hz EMA, the question remained regarding if higher RT-MRI frame rates, along with a larger dataset, could have a positive impact on the outcomes. Since one frame is used as representative of the vocal tract configuration for each sound, a higher frame rate, might enable capturing key moments of articulation, e.g., alveolar contact for the /n/.

At the onset of MEMNON, and to address these aspects, the authors [26] explored the novel 50Hz RT-MRI data showing both the applicability of the methods along with an improvement of the results. At this point, the vocal tract configurations serving as input for the method were defined as sets of landmarks placed on the lips, tongue surface, and velum, to gather data similar to what was possible with EMA, mimicking the original application of the method in order to establish its applicability considering the novel data. However, the data available from RT-MRI can potentiate the consideration of different tract variables beyond landmarks. In this regard, the authors have explored tract variables aligned with the Task Dynamics framework [27], adopting the concept of constrictions (except for the velum), showing them as an alternative that was more compact (less variables involved) and provided interesting critical articulator results with the advantage of a more direct relation with existing Articulatory Phonology descriptions and, hence, fostering an easier discussion based on the existing literature. In the most recent work, in this regard, the team explored a novel representation for the velum, to completely avoid landmarks, and considering both the orovelar and velopharyngeal passages [28].

4.3.2. Analysis of EP nasals dynamics

Regarding the oral configuration of EP nasal vowels, we started by exploring a new quantitative method to systematically evaluate articulatory information [29]. The method provides new insights of the oral adjustments in nasal vowels resorting to constriction location and degree as variables to describe the vocal tract configuration aligned with the Articulatory Phonology framework.

More recently, and taking into consideration a set of methods proposed by Carigan et al. [30], the RT-MRI data of 11 speakers was processed to explore how generalized additive mixed models (GAMMs) and functional linear mixed models (FLMMs) might contribute to the analysis of nasal vowel dynamics and coordination. In this regard, GAAMs were computed for the tract aperture function, over time, and FLMMs were computed for tract apertures at 20%, 50% and 80% of the vowel interval to provide a more detailed analysis of vocal tract configuration at specific time points (beginning, middle and end of the vowel).

Additionally, FLMMs were also used to model lips and velum aperture over time. Figure 2 presents an example of the

FLMM of lips and velum for the vowel [a], and a detailed report of these results has also been submitted to this conference [31].

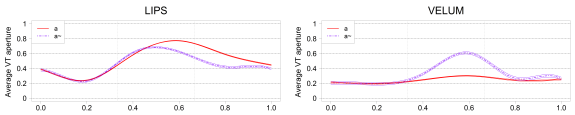


Figure 2: *Lip and velar aperture, over time, for vowels [a], in red, and [ẽ], in blue, obtained by applying FLMMs to the data obtained from 11 EP speakers. The plot also shows the data corresponding to the flanking consonants [p] (before 0.33) and [t] (after 0.66).*

4.4. Audiovisual Speech Synthesis

The work regarding audiovisual speech synthesis considering an articulatory approache evolves previous work performed in Marie Curie IAPP project IRIS [32] and is grounded on the vision described by our team to advance the field [24].

Continuing the previous work on articulatory audio-visual synthesis (AVS) system, several improvements were made. Figure 3 A) presents the complete architecture of the system, with all the modules. In summary, the system receives, as input, a sentence in European Portuguese, converts it from grapheme to phoneme (G2P), applies syllabification, and this is injected in the Task Dynamics model to compute articulator trajectories. These, are then used to manipulate rigs in a photorealistic 3D model. The most significant updates, in the scope of MEMNON, were the creation of two new modules supporting G2P and syllabification. The implementation of these modules uses machine learning methods to process the data.

One of aspects that is being addressed concerns the quality of the output audio that is currently provided by an articulatory synthesizer. The goal is to profit from the advantages of having an articulatory approach (e.g., with individual control over the articulators), but improve the audio quality. To this end, some exploration is currently ongoing with WaveGlow [33], trying to improve the speech synthesis module.

To foster easier integration of this audiovisual synthesis system in applications, e.g., for evaluation, it was developed a first prototype interaction modality, aligned with the W3C recommendation for multimodal interaction systems and supporting multi-device interaction [34]. The modality is composed of different modules: 1) on the user side runs a part of the modality, the viewer; 2) in a cloud server runs the other part of the modality were the avatar is generated, also this part communicates with the modality support services and the Interaction Manager. The overall architecture of the multimodal system and modality is presented in Figure 3B) and in Figure 3C) a preview of the interface of the modality showing the AVS stream.

4.5. Silent Speech Interfaces

The work on silent speech is also ongoing and the team expects to have first results of exploring noninvasive technologies in the near future.

5. Conclusions

This paper presents research project MEMNON, focusing on the motivation and challenges considered at its genesis, the adopted approach, and highlights of already achieved results. This overall report aims at both dissemination of the methods

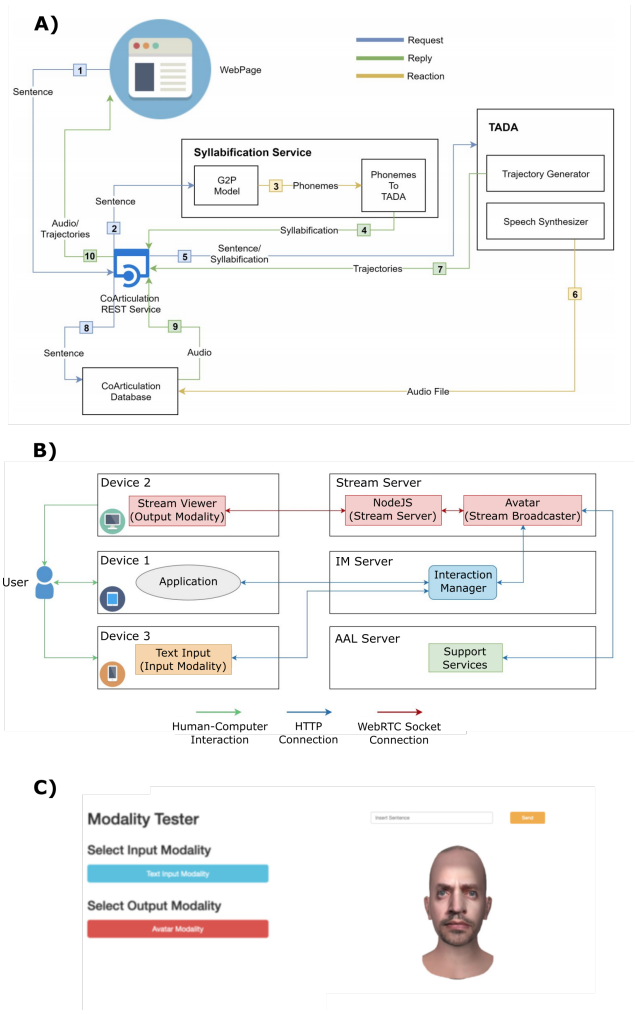


Figure 3: *A) Audio-Visual Speech system architecture; B) Architecture of the modality in a multimodal system; B) Preview of the AVS modality*

and results and an evangelization for the relevance of speech production studies.

At this point, we have started dynamic analysis of the RT-MRI data with first results regarding oral and nasal vowel coordination and are advancing an online platform to support the speech production studies workflow.

Currently, the line of work deserving the strongest research efforts is silent speech interfaces exploring non-invasive technologies in line with previous efforts of the team, e.g., with ultrasonic Doppler. The outcomes of this effort should inform the proposal of novel interactive technologies suitable, for instance, in the scope of recent AAL Programme project APH-ALARM in which our team is collaborating to propose assistive technologies for aphasia scenarios.

6. Acknowledgements

This work is partially funded by the German Federal Ministry of Education and Research (BMBF, with the project 'Synchonic variability and change in European Portuguese'), by IEETA Research Unit funding (UIDB/00127/2020), by Portugal 2020 under COMPETE Program , and the European Regional Development Fund through project SOCA – Smart Open Campus (CENTRO-01-0145-FEDER-000010), and project MEMNON (POCI-01-0145-FEDER-028976).

7. References

[1] A. D. Scott, R. Boubertakh, M. J. Birch, and M. E. Miquel, "Adaptive averaging applied to dynamic imaging of the soft palate," *Magnetic resonance in medicine*, vol. 70, no. 3, pp. 865–874, 2013.

[2] S. Silva and A. Teixeira, "Critical articulators identification from RT-MRI of the vocal tract," in *Proc. Interspeech 2017*, Stocholm, Sweden, August 2017.

[3] P. Martins, I. Carbone, A. Silva, and A. Teixeira, "Coarticulatory effects on european portuguese: A first MRI study," in *La co-articulation: Indices, Direction et Représentation, Workshop de l'Association Francophone de la Communication Parlée*. DIPRALNAG et PRAXILING, Nov. 2007.

[4] A. Teixeira, C. Oliveira, and P. Barbosa, "European portuguese articulatory based text-to-speech: First results," in *Computational Processing of the Portuguese Language, The International Conference on Computational Processing of Portuguese, PROPOR 2008, Lecture Notes in Computer Science/LNAI, Vol. 5190*. Springer, 2008.

[5] S. Silva and A. Teixeira, "Unsupervised segmentation of the vocal tract from real-time mri sequences," *Computer Speech and Language*, 2015, [ISI ARTICLE].

[6] ——, "Quantitative systematic analysis of vocal tract data," *Computer Speech and Language*, vol. 36, pp. 307–329, March 2016.

[7] A. Teixeira and F. Vaz, "European Portuguese Nasal Vowels: An EMMA study," in *7th European Conference on Speech Communication and Technology, EuroSpeech - Scandinavia*, vol. 2. Aalborg, Dinamarca: CPK/ISCA, Sep. 2001, pp. 1843–1846.

[8] P. Martins, I. Carbone, A. Silva, and A. Teixeira, "An MRI study of European Portuguese nasals," in *Interspeech*, 2007.

[9] S. Rossato, A. Teixeira, and L. Ferreira, "Les nasales du Portugais et du Français : une étude comparative sur les données EMMA," in *JEP'2006*, Rennes, França, 2006.

[10] P. Martins, I. Domingues, A. Silva, and A. Teixeira, "Effects coarticulatoires sur le portugais européen: une première étude IRM," in *Production et Perception de la Parole : La coarticulation, des Indices aux Représentations.*, ser. Collection Langue & Parole, M. EMBARKI and C. DODANE, Eds. L'Harmattan, 2012.

[11] S. Silva, P. Martins, C. Oliveira, and A. Teixeira, "Quantitative analysis of /l/ production from RT-MRI: First results," in *Advances in Speech and Language Technologies for Iberian Languages*, ser. LNAI. Springer, November 2014.

[12] J. Freitas, A. Teixeira, M. S. Dias, and C. A. C. Bastos, "Towards a multimodal silent speech interface for european portuguese," in *Speech Technologies*. INTECH, 2011.

[13] J. Freitas, M. S. Dias, and A. Teixeira., "Towards a silent speech interface for portuguese: Surface electromyography and the nasality challenge." in *International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS 2012),*, Vilamoura, Portugal, Feb. 2012.

[14] J. Freitas, A. Teixeira, F. Vaz, and M. S. Dias, "Automatic speech recognition based on ultrasonic doppler sensing for european portuguese," in *Advances in Speech and Language Technologies for Iberian Languages, vol. CCIS 328, Springer*, 2012.

[15] J. Freitas, A. Teixeira, and M. S. Dias., "Multimodal silent speech interface based on video, depth, surface electromyography and ultrasonic doppler: Data collection and first recognition results," in *Workshop on Speech Production in Automatic Speech Recognition*, Lyon, aug 2013.

[16] J. Freitas, S. Silva, A. Teixeira, and M. S. Dias, "Assessing the applicability of surface emg to tongue gesture detection," in *Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2014.

[17] S. Silva and A. Teixeira, "An anthropomorphic perspective for audiovisual speech synthesis," in *Proc. BIOSIGNALS*, Feb. 2017.

[18] J. Freitas, A. Teixeira, M. S. Dias, and S. Silva, *An Introduction to Silent Speech Interfaces*. Springer, July 2016.

[19] J. Freitas, S. Candeias, M. S. Dias, E. Lleida, A. Ortega, A. Teixeira, S. Silva, C. Acarturk, and V. Orvalho, "The IRIS Project: A liaison between industry and academia towards natural multimodal communication," in *Proc. IberSpeech 2014*, 2014.

[20] M. Uecker, S. Zhang, D. Voit, A. Karaus, K.-D. Merboldt, and J. Frahm, "Real-time mri at a resolution of 20 ms," *NMR in Biomedicine*, vol. 23, no. 8, pp. 986–994, 2010.

[21] N. Almeida, S. Silva, A. Teixeira, and C. Cunha, "Collaborative quantitative analysis of rt-mri," in *Proc. 12th International Seminar on Speech Production (ISSP), (accepted)*, 2020.

[22] S. Silva and A. Teixeira, "Unsupervised segmentation of the vocal tract from real-time MRI sequences," *Computer Speech and Language*, vol. 33, no. 1, pp. 25–46, Sep. 2015.

[23] C. Oliveira, "Do grafema ao gesto - contributos linguísticos para um sistema de síntese de base articulatória," PhD Thesis, Universidade de Aveiro, March 2009.

[24] S. Silva and A. J. Teixeira, "An anthropomorphic perspective for audiovisual speech synthesis." in *BIOSIGNALS*, 2017, pp. 163–172.

[25] P. J. Jackson and V. D. Singampalli, "Statistical identification of articulation constraints in the production of speech," *Speech Communication*, vol. 51, no. 8, pp. 695 – 710, 2009.

[26] S. Silva, A. Teixeira, C. Cunha, N. Almeida, A. A. Joseph, and J. Frahm, "Exploring Critical Articulator Identification from 50Hz RT-MRI Data of the Vocal Tract," in *Proc. Interspeech 2019*, 2019, pp. 874–878. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2897

[27] S. Silva, C. Cunha, A. Teixeira, A. Joseph, and J. Frahm, "Towards automatic determination of critical gestures for european portuguese sounds," in *Proc. International Conference on Computational Processing of the Portuguese Language (PROPOR)*. Springer, 2020, pp. 3–12.

[28] S. Silva, N. Almeida, C. Cunha, A. Joseph, J. Frahm, and A. Teixeira, "Data-driven critical tract variable determination for European Portuguese," *Information*, vol. 11, no. 10, p. 491, 2020.

[29] C. Cunha, S. Silva, A. Teixeira, C. Oliveira, P. Martins, A. A. Joseph, and J. Frahm, "On the Role of Oral Configurations in European Portuguese Nasal Vowels," in *Proc. Interspeech 2019*, 2019, pp. 3332–3336. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2232

[30] C. Carignan, P. Hoole, E. Kunay, M. Pouplier, A. Joseph, D. Voit, J. Frahm, and J. Harrington, "Analyzing speech in both time and space: Generalized additive mixed models can uncover systematic patterns of variation in vocal tract shape in real-time mri," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 11, no. 1, 2020.

[31] N. Almeida, C. Cunha, , S. Silva, and A. Teixeira, "Data-driven analysis of nasal vowels dynamics and coordination in bilabial contexts," in *[submitted to IberSpeech 2020]*, 2020, (submitted).

[32] S. Silva, A. Teixeira, and V. Orvalho, "Articulatory-based audiovisual speech synthesis: Proof of concept for European Portuguese," in *Proc. Iberspeech*, Lisbon, Portugal, 2016, pp. 119–126.

[33] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[34] N. Almeida, S. Silva, A. Teixeira, and D. Vieira, "Multi-Device Applications Using the Multimodal Architecture," in *Multimodal Interaction with W3C Standards - Toward Natural User*, 1st ed., D. Dahl, Ed. Springer International Publishing, 2017, pp. 367–383.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Towards conversational technology to promote, monitor and protect mental health

Zoraida Callejas[1], David Griol[1], Kawtar Benghazi[1], Manuel Noguera[1], María Inés Torres[2], Raquel Justo[2], Anna Esposito[3], Gennaro Cordasco[3], Raymond Bond[4], Maurice Mulvenna[4], Edel Ennis[4], Siobhan O'Neill[4], Huiru Zheng[4], Matthias Kraus[5], Nicolas Wagner[5], Wolfgang Minker[5], Gavin McConvey[6], Matthias Hemmje[7], Michael Fuchs[7], Neil Glackin[8], Gérard Chollet[8]

[1]University of Granada, Granada, Spain,
[2]University of the Basque Country, Bilbao, Spain
[3]Università degli Studi della Campania "Luigi Vanvitelli", Casserta, Italy
[4]Ulster University, Belfast, Northern Ireland
[5]Ulm University, Ulm, Germany
[6]Action Mental Health, Newtonards, Northern Ireland
[7]GLOBIT, Pfungstadt, Germany
[8]Intelligent Voice, London, United Kingdom

[1]{zoraida, dgriol, benghazi, mnoguera}@ugr.es,
[2]{manes.torres, raquel.justo}@ehu.eus,
[3]iiass.annaesp@tin.it, gennaro.cordasco@unicampania.it
[4]{rb.bond, md.mulvenna, e.ennis, sm.oneill, h.zheng}@ulster.ac.uk,
[5]{matthias.kraus, nicolas.wagner, wolfgang.minker}@uni-ulm.de,
[6]gmcconvey@amh.org.uk,
[7]{Matthias.Hemmje, Michael.Fuchs}@globit.com ,
[8]{neil.glackin, gerard.chollet}@intelligentvoice.com

Abstract

This paper presents a general overview of the H2020-MSCA-RISE project MENHIR (Mental health monitoring through interactive conversations), which aim is to explore the possibilities of conversational technologies (chatbots) to understand, promote and protect mental health and assist people with anxiety and mild depression manage their conditions. MENHIR started on February 2019 and will have a duration of 4 years. Its consortium brings together 8 partners including universities, a non-profit organization and companies.

Index Terms: mental e-health, conversational systems, chatbots, dialogue systems, wellbeing

1. Introduction

Mental wellbeing is a very important aspect of health. New technology solutions such as chatbots are potential channels for supporting and coaching users to maintain a healthy lifestyle [1, 2] and in particular a good state of mental wellbeing. Conversational systems[1] have the added value of providing social conversations and coaching 24/7 outside from conventional mental health services.

The MENHIR project (http://menhir-project.eu) has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823907, concretely in the funding scheme Marie Sklodowska-Curie Research and Innovation Staff Exchange (MSCA-RISE) 2018.

MENHIR aims at engaging people in the management of self-care and providing a more flexible, immediate and sustained complement to under pressure mental health services [3]. The MENHIR chatbot technology provides symptom and mood management, personalized support and motivation, coping strategies, mental health education, signposting to online resources and complementing the support received by local services.

The technology developed will be able to process emotional and mood patterns from different aspects of the voice recorded (e.g. from the tone of voice or speed of speech). This information facilitates rapid intervention and appropriate feedback such as prompts and supportive messages. It will also generate personalized conversations with the users to motivate and engage them at any time and especially in between visits to their local services.

The limitations of previous technology are the high dropout rates due in part to a lack of interactive features and appropriate prompts, difficulties for people with low levels of technological literacy and failure to provide human support. The MENHIR chatbot technology aims to address these difficulties being more usable and intuitive since it simulates everyday human to human conversation.

The project will also generate data on how people with anxiety and depression interact with and respond to this technology, and about the benefits it brings to them, which will be useful to understand the role that chatbots can play to promote mental health.

[1]We will refer to them as chatbots. Although chatbots are frequently used to describe conversational systems with which to maintain open chit-chat (non task-oriented) conversations, we use the term as a synonym of a general conversational system or dialogue system. We have adopted this term because it is more recognizable by the end users who are co-creating the technology in MENHIR.

10.21437/IberSPEECH.2021-32

2. Consortium

The MENHIR consortium is comprised of eight partners:

- University of Granada, Spain (coordinator, academic).
- University of the Basque Country, Spain (academic).
- Università degli Studi della Campania "Luigi Vanvitelli", Italy (academic).
- Ulster University, United Kingdom (academic).
- Ulm University, Germany (academic).
- Action Mental Health, United Kingdom (non-profit).
- GLOBIT - Globale Informationstechnik GmbH, Germany (company).
- Intelligent Voice Limited, United Kingdom (company).

3. Objectives

MENHIR has been funded by the RISE programme, which aim is the sharing of knowledge and ideas across countries, sectors and disciplines. Thus, one of the main objectives of the project is to develop new capabilities and skills for MENHIR participants.

The staff participating in MENHIR will create new opportunities for IT researchers to increase their impact rooting technology in a deeper understanding of their prospective users, while psychologists and mental health experts will gain knowledge of the technological tools that can help them improve their clients' lives.

The main scientific objectives are to: i) establish the strengths, limitations and requirements of mental health chatbots; ii) develop the MENHIR chatbot technology collaborating with people who suffer anxiety and/or mild depression; and iii) understand how users engage with chatbot technology and how it may affect their well-being over time.

4. Progress beyond the state of the art and potential impacts

MENHIR will progress beyond the state of the art by:

- Determining the most practical scales for monitoring mental wellbeing.
- Compiling mental health coping strategies, prompts, relevant digital content and resources.
- Providing new evidence on how to perform a cross-modal analysis of the interactional exchanges to automatically recognize the users' emotional state and anxiety level.
- Defining how to track mood and anxiety from user-system interactions and how to represent user progress in a computational model.
- Providing novel approaches to generate computerized models of the users to adapt decision making, conversational models and communication styles.
- Finding novel approaches to make chatbot interaction personalized to their users, taking into consideration that they are heterogeneous in terms of age, gender, mental health condition, progression and responses to system's strategies.
- Identifying adequate dialogue management strategies based on the user progress and the history of the user-system interaction.
- Establishing adequate proactivity approaches and disparate conversation structures to manage conversation and favour user self-disclosure.
- Establishing new methods to favour user-system trust through active listening.
- Generating good practices for data storage, annotation and sharing.
- Generating good practices for participatory research based on co-creation.

These results are being documented in several reports that are being delivered during the project, corresponding to: the chatbot strategy for mental health monitoring, the characterization and analysis of human-to-human interaction, the cross-modal analysis and annotations of mood and anxiety levels, user modelling (including mood and anxiety level), and conversation modelling.

MENHIR will have a considerable impact, as it presents a unique research and innovation plan which tackles several challenges from different areas merging multiple expertise of academic and non-academic members with a background in Psychology, Cognitive Sciences, Computer Science and Engineering. MENHIR will help to achieve the vision that should be accomplished by 2025 according to the Strategic Roadmap for the area of multimodal conversational interaction technologies [4], as it will help to achieve several of the multiple goals defined in the roadmap, including:

- Multimodal conversational interfaces with the potential to adapt automatically to the user and to the user's state.
- Make predictions and recommendations based on personal data with a sensitive management of information.
- Domain-specific or context-based personal assistants.
- Dialogue modelling and management with cognitive human modelling, flexible domain adaptation and socially acceptable dialogues.
- Ability to understand emotions, sentiment and social signals as well as the meaning of the spoken word.

5. Recent advances and current work

MENHIR chatbot technology is a complement to the efforts of mental health service providers, so it was important that all MENHIR members became aware of the ways in which technology can be useful to them and their clients and avoid introducing functionality that could be unintentionally detrimental for the user's wellbeing. This was achieved during mutual visits (secondments).

In MENHIR we have adopted a co-creation methodology. This means that the prospective users participate in the design and development of technology not as passive recipients, but as active members that are engaged from the beginning in the relevant decisions. Prospective users are represented in MENHIR by clients of Action Mental Health (AMH) across its different services in Northern Ireland. A co-creation workshop was performed on June 2019. The participants were 9 AMH clients, a key worker and 3 MENHIR researchers who presented the project and acted as facilitators. This was a lively and very interesting conversation that provided relevant insight for the project.

Among the results of the workshop, we elaborated a list of the perceived strengths and limitations of the technology and

also a study of the main topics and themes covered, which included the challenges faced by people such as isolation and difficulty for honest disclosure, symptom recognition, continuous monitoring, disclosure facilitation, companionship, risk detection, personalization, configurable proactiveness, user access to information, privacy, vulnerability, cost and access to the technology and use of the chatbot. The results are the co-created functional requirements and a number of use case scenarios that can be of interest to guide future development of chatbots in the mental health domain are described in further detail in [5].

The results of the co-creation together with a detailed analysis of the scientific literature about mental health monitoring systems and multimodal emotion and mood recognition, allowed to define the scenarios in which it will be used, which will be mainly four: intelligent reminder, diary, progress evaluation and risk detection.

In the intelligent reminder scenario, the chatbot will remind users to engage in the activities planned for them. The reminder will be capable of making adequate decisions on how, when and if at all to remind and motivate users to stick to their plans without being obtrusive. This will be achieved by means of a rich user model based on the user's objectives and engagement, challenges that are pervasive to different types of multimodal e-coaches [6]. Adaptive, trustworthy dialogue management will be key to this task [7, 8], for which different models will be put in place an evolved as more data is available [9, 10]. In relation to this, we are conducting several studies related to the acceptance of different aspects of the interaction with a conversational system in multiple settings. [11, 12].

The diary use case provides users the possibility to talk to the system about their day. The chatbot provides follow-up questions to build rapport and keep the user engaged and talking. The goal is to offer companionship and record the user interventions. Then, we will develop technology that will be able to infer the user's emotional state from the recordings using paralinguistic information considering the different input sources and interaction settings [13].

In MENHIR we are devoting a great effort to data collection, annotation, representation and sharing, see e.g. [14]. We have performed and annotated anonymized recordings of AMH clients and a control group that we are currently processing to obtain relevant features for user state recognition.

The progress evaluation use case will be based on the adaptation of questionnaires designed to assess mental health. This serves to address the need for continuous monitoring using standardized and widely recognized questionnaires, and will be used to check whether the state recognized automatically from the recorded entries to the user diary is reliable.

Risk detection functionalities are pervasive to the previous scenarios and consist in detecting and notifying risks based on predefined words and phrases and raise an alert.

6. Acknowledgements

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823907 (MENHIR: Mental health monitoring through interactive conversations `https://menhir-project.eu`).

7. References

[1] A. Benítez-Guijarro, Z. Callejas, M. Noguera, and K. Benghazi, "Architecting dietary intake monitoring as a service combining NLP and IoT," *Journal of Ambient Intelligence and Humanized Computing*, 2019.

[2] D. Griol and Z. Callejas, "Mobile Conversational Agents for Stroke Rehabilitation Therapy," in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, 2019, pp. 513–518.

[3] Z. Callejas, K. Benghazi, M. Noguera, M. I. Torres, and R. Justo, "MENHIR: Mental health monitoring through interactive conversations," *Procesamiento del Lenguaje Natural*, vol. 63, 2019.

[4] S. Renals, J. Carletta, K. Edwards, H. Bourlard, P. Garner, A. Popescu-Belis, D. Klakow, A. Girenko, V. Petukova, P. Wacker, A. Joscelyne, C. Kompis, S. Aliwell, W. Stevens, and Y. Sabbah, "ROCKIT: Roadmap for Conversational Interaction Technologies," in *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges*, ser. RFMIR '14. New York, NY, USA: Association for Computing Machinery, 2014, pp. 39–42.

[5] A. Benítez-Guijarro, R. Bond, F. Booth, Z. Callejas, E. Ennis, A. Esposito, M. Kraus, G. McConvey, M. McTear, M. Mulvenna, C. Potts, L. Pragst, R. Turkington, N. Wagner, and H. Zheng, "Co-creating Requirements and Assessing End-User Acceptability of a Voice-Based Chatbot to Support Mental Health: A Thematic Analysis of a Living Lab Workshop," in *Conversational Dialogue Systems for the Next Decade*, L. F. D'Haro, Z. Callejas, and S. Nakamura, Eds. Singapore: Springer, 2021, pp. 201–212.

[6] L. Angelini, M. El Kamali, E. Mugellini, O. Abou Khaled, Y. Dimitrov, V. Veleva, Z. Gospodinova, N. Miteva, R. Wheeler, Z. Callejas, D. Griol, K. Benghazi, M. Noguera, P. Bamidis, E. Konstantinidis, D. Petsani, A. Beristain Iraola, D. I. Fotiadis, G. Chollet, I. Torres, A. Esposito, and H. Schlieter, "First Workshop on Multimodal e-Coaches," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, ser. ICMI '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 890–892.

[7] D. Griol, Z. Callejas, J. M. Molina, and A. Sanchis, "Adaptive dialogue management using intent clustering and fuzzy rules," *Expert Systems*, vol. 38, no. 1, 2021.

[8] M. Kraus, P. Seldschopf, and W. Minker, "Towards the Development of a Trustworthy Chatbot for Mental Health Applications," in *27th International Conference on Multimedia Modeling*. Prague, Czech Republic: Springer, 2021.

[9] D. Griol and Z. Callejas, "Discovering Dialog Rules by Means of an Evolutionary Approach," in *Interspeech 2019*. ISCA, 2019, pp. 1473–1477.

[10] A. Esposito, M. Faundez-Zanuy, F. C. Morabito, and E. Pasero, "Some Note on Artificial Intelligence," in *Neural Approaches to Dynamics of Signal Exchanges*, ser. Smart Innovation, Systems and Technologies. Singapore: Springer, 2020, pp. 3–8.

[11] A. Esposito, T. Amorese, M. Cuciniello, A. M. Esposito, A. Troncone, M. I. Torres, S. Schlögl, and G. Cordasco, "Seniors' Acceptance of Virtual Humanoid Agents," in *Ambient Assisted Living*. Springer, Cham, 2018, pp. 429–443.

[12] A. Esposito, T. Amorese, M. Cuciniello, M. T. Riviello, A. M. Esposito, A. Troncone, and G. Cordasco, "The Dependability of Voice on Elders' Acceptance of Humanoid Agents," in *Interspeech 2019*. ISCA, 2019, pp. 31–35.

[13] A. Benítez-Guijarro, Z. Callejas, M. Noguera, and K. Benghazi, "Coordination of Speech Recognition Devices in Intelligent Environments with Multiple Responsive Devices," *Proceedings*, vol. 31, no. 1, p. 54, 2019.

[14] B. Vu, M. deVelasco, P. Mc Kevitt, R. Bond, R. Turkington, F. Booth, M. Mulvenna, M. Fuchs, and M. Hemmje, "A Content and Knowledge Management System Supporting Emotion Detection from Speech," in *Conversational Dialogue Systems for the Next Decade*, ser. Lecture Notes in Electrical Engineering, L. F. D'Haro, Z. Callejas, and S. Nakamura, Eds. Singapore: Springer, 2021, pp. 369–378.

GENIOVOX Project: Computational generation of expressive voice

Oriol Guasch, Francesc Alías, Marc Arnela, Joan Claudi Socoró, Marc Freixes and Arnau Pont*

GTM-Grup de Recerca en Tecnologies Mèdia, La Salle, Universitat Ramon Llull
C/Quatre Camins 30, 08022 Barcelona, Catalunya, España.

{oriol.guasch, *francesc.alias, marc.arnela, joanclaudi.socoro, marc.freixes, arnau.pont}@salle.url.edu

Abstract

The GENIOVOX project: "Computational synthesis of expressive voice", with ref. TEC2016-81107-P and funded by the Ministerio de Economía, Industria y Competitividad (Plan Nacional de I+D Excelencia) was carried out in the period 2016-2019. Its two main objectives were the following ones. On the one hand, diphthongs and hiatuses were simulated in three-dimensional (3D) geometries using the finite element method (FEM), based on the resolution of the underlying wave equations. Likewise, techniques were developed to simulate syllables with fricative consonants that did not require the use of high-performance computing. The trick was to approximate the interdental flow acoustic source terms using quadrupole, dipole and monopole distributions instead of getting them from a computational fluid dynamics simulation. In addition to generating diphthongs and syllables with fricatives, the project proposed a first attempt to incorporate some expressive effects through modifications of the vocal tract geometry and the glottal source model. Vowel sounds were computationally generated by convoluting the impulse response of 3D FEM vocal tracts with glottal pulses that incorporated tense, neutral and lax phonations from expressive speech corpora.

1. Introduction

Pronouncing a sound as simple as a vowel is extremely easy for us. However, in doing so we are unaware of the large number of physical phenomena involved. The turbulent flow of air exhaled from the lungs induces self-oscillations of the vocal cords. These generate acoustic waves that propagate inside the vocal tract and are emitted outwards. As we vary the shape of the vocal tract (VT) we will perceive one sound or another. At present, the numerical simulation of voice from the modeling of its underlying physics by means of realistic three-dimensional (3D) geometries of the VT is a notable challenge, making it necessary to resort to supercomputing centers to simulate certain sounds. For example, it is not possible to produce a fricative using the finite element method (FEM) with a desktop computer, although we can synthesize a vowel or a diphthong, with good equipment. On the other hand, so far, numerically generated sounds have lacked expressiveness. This aspect is not only important to correctly emulate communication between people, but it can also be related to certain medical issues. For instance, it is possible to detect pathological aspects of the voice from the sustained pronunciation of a simple vowel, or from a diphthong.

The two main contributions of the GENIOVOX project are as follows. To begin with, computational simulations of diphthongs and hiatuses have been carried out, enriching the number and type of computational sounds generated to date with dynamic 3D VTs. These sounds have been generated through interpolation between static vowel VT geometries and, in a more realistic way, by resorting to the biomechanical model ArtiSynth (https://www.artisynth.org/). More importantly, methods for synthesizing fricatives in 3D geometries that do not require fluid dynamics calculations in supercomputing centers have been developed. Aerodynamic sources of noise have been simulated using quadrupole and/or dipole and monopole source distributions, inspired to some extent by 1D-based approaches. This has allowed us to generate syllables and sequences such as /sa/ and /asa/ for 3D VTs with desktop computers and at a cost not much higher than that of generating diphthongs.

The second major contribution of the GENIOVOX project is that, for the first time, the possibility of including some expressive effects on FEM generated voice has been attempted. Such effects can be achieved from changes in phonation and/or in the VT geometry. Both aspects have been taken into account. On the one hand, the tense, modal and lax voice continuum in the FEM generation of vowels has been analyzed using simplified and realistic 3D geometries of the vocal tract. Also, numerical methods have started being developed to modify the geometry of the vocal tract in order to achieve effects such as the grouping of the formants of a vowel (i.e. the so-called singing formant), an important element, for example, to simulate the projection of the sung voice.

2. Goals and main results

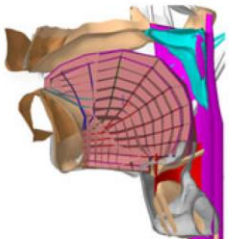
The six main objectives of the GENIOVOX project and the results achieved during its execution are summarized below.

2.1. O1: Characterization of expressive parameters in recorded voice

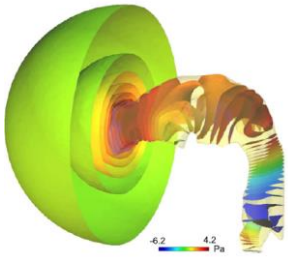
The GTM has five voice corpora corresponding to the following expressive styles: neutral, happy, aggressive, sad and sensual. Moreover, we analyzed a storytelling corpus containing diverse expressive categories. The analyses focused on the set of words that appear in all corpora, and more specifically, on the vowels. Prosodic parameters (fundamental frequency, duration and energy), perturbation parameters (jitter and shimmer) and several voice quality parameters (e.g., harmonic-to-noise ratio, Hammarberg index, spectral slope, etc.) were extracted to characterize the main components of expressive speech.

On the other hand, a proposal was developed to add singing capabilities to a neutral corpus-based text-to-speech synthesis system. A text-to-speech-and-singing synthesis framework that integrates speech-to-singing conversion was developed and tested, achieving reasonable quality compared to corpus-based singing approaches to eventually address singing needs in storytelling applications according to the conducted MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) perceptual test.

10.21437/IberSPEECH.2021-33



a)



b)

Figure 1: *a) Obtaining the vocal tract for the vowel /a/ from the ArtiSynth biomechanical model. b) Propagation of acoustic waves inside the vocal tract of an /a/, obtained with FEM. Image published in [5].*

2.2. O2: Simulation of diphthongs / hiatuses

This objective focused on the simulation of the diphthongs /ai/, /au/ and /ui/ in 3D dynamic geometries of the VT by means of FEM. This requires solving the wave equation in a domain that evolves from the geometry of the initial vowel to that of the final vowel. Two different strategies were developed for this purpose. First, static geometries generated from MRI (Magnetic Resonance Imaging) were discretized using an adaptive grid and a semi-polar grid, which allowed us to simplify the interpolation between 3D meshes. Second, a methodology was developed to extract closed cavities from the ArtiSynth biomechanical model and thus unify, for the first time, biomechanical and acoustic simulations in three dimensions. All the acoustic simulations were carried out by means of an in-house developed FEM code (see Figure 1).

2.3. O3: Simulation of syllables with fricative consonants

In this objective, the fricatives /s/ and /z/ were simulated with FEM. On the one hand, some computational aeroacoustics (CAA) simulations that required the use of supercomputing were completed. On the other hand, and given that precisely the computational cost of generating a syllable that contains a fricative is prohibitive, even in a supercomputing center, a method was developed to avoid the computational fluid dynamics (CFD) calculation of the simulation, which allows treating only with the wave equation. The contribution consists in approximating the source term that we would obtain from the CFD by means of a random distribution of Kirchhoff vortices (see Figure 2). In this way, the cost of synthesizing a syllable like /sa/ is not excessively higher than that of generating a diphthong. As a third alternative, monopolar and dipole sources excited using Gaussian noise were implemented. This allowed us to synthesize the sequence /asa/ with simplified 3D geometries.

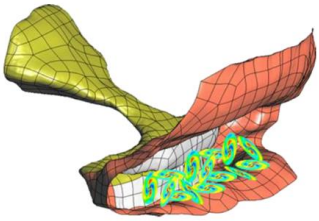


Figure 2: *Key idea for the generation of fricatives without CFD calculation: approximation of the flow source term by means of a random distribution of Kirchhoff's vortices.*

2.4. O4: Variation and adjustment in time of the geometry of the vocal tract

Varying the shape of the vocal tract to, say, concentrate formants and thus achieve certain expressive effects on the voice, is usually carried out in the context of 1D articulatory synthesis using sensitivity functions. These are justified from the non-linear phenomenon of radiation pressure. Throughout the project, some theoretical advances in this line were carried out. In particular, it was shown that the sensitivity functions can be obtained from a perturbation analysis of the VT eigenmodes, without the need to resort to the non-linear phenomenon of radiation pressure to justify them. This totally different theoretical approach is more prone to updating techniques in FEM and will facilitate optimizing 3D VT shapes to achieve expressive effects in the future.

2.5. O5: Glottal pulse modification

Throughout the project, the LF (Liljencrants-Fant) glottal pulse model was used to generate the excitation signal. This parametric model allowed us to modify the shape of the pulses using the control parameter Rd, in order to simulate lax, modal and tense phonations. To that goal, an aliasing-free LF model was adapted incorporating the control parameter Rd. Furthermore, this implementation was extended by integrating an aspiration model to emulate the aspiration noise found in vowels. This allowed us to obtain more realistic glottal pulses and generate more natural voice. The modification of the glottal pulses using the herein detailed method was evaluated in the generation of different vowels. The frequency responses obtained from a realistic geometry and a simplified one by FEM were used. The evaluation was made from the spectral analysis of the results and by quantifying the high frequency energy content.

2.6. O6: Numerical generation of expressive voice

The first steps were taken for the inclusion of expressive effects in the generation of computational voice. Specifically, simplified and realistic 3D vowel tracts were considered for the vowels /a/, /i/ and /u/. The former were based on models with radial symmetry while the latter consisted of realistic MRI-based models. The impulse responses of the vocal tracts were calculated by FEM and then convoluted with the glottal pulses corresponding to the tense, neutral and lax phonations (see Figure 3). From there, the influence that the type of phonation has on the high-frequency energy content of the different vowels according to the considered geometries was analyzed.

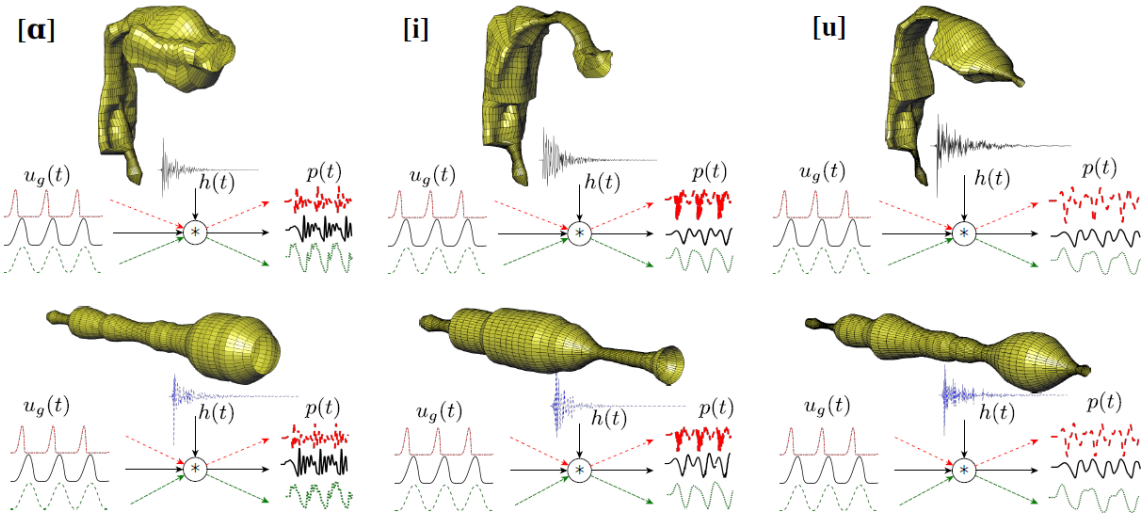


Figure 3: *Synthesis of the vowels /a/, /i/ and /u/ with realistic and simplified vowel tracts. The acoustic signal p(t) is calculated from the convolution of the glottal source $u_g(t)$ with the impulse response h(t) obtained from the 3D finite element model. Three types of phonation are considered: Tense (dashed red line), modal (black line) and lax (dashed green line). Image published in [6].*

2.7. O7: Evaluation of the quality of the generated voice

From the voice corpus analyzed in O1, the three corresponding to modal and tense voice were chosen from neutral, happy and aggressive styles, respectively. Parallel realizations in these three expressive styles were analyzed using a glottal vocoder that decomposes the speech signal into glottal excitation and VT response, and independently parameterizes them. From these parameters, two expressive conversion models (LF-based glottal excitation and VT) were trained. The neutral words were converted to the happy and aggressive styles using expressive prosody and, through different configurations of the conversion models, the contribution of the spectral characteristics of arousal and the vocal tract in the generation of a happy and aggressive voice was studied. The objective evaluation focused on the spectral analysis of the results and on the calculation of spectral distances between each configuration and the expressive reference on vowels using carrier words. The subjective evaluation was carried out using a MUSHRA perceptual test.

3. Publications

As a result of the project, 11 articles were published in journals from different fields (from numerical methods to acoustics and speech processing). Likewise, 13 papers were presented at several international conferences such as InterSpeech2017, ISSP2017, ECCM-ECFD2018, IberSPEECH2018, Inter-Noise2019, ECCOMAS-YIC2019, ICA2019, InterSpeech2019 and SSW10, and 1 book chapter was also published. The list of all publications is presented in the References' section of this article.

Finally, it is also worth mentioning that the PhD Thesis of Marc Freixes has been partially developed within the GENIOVOX project, under the supervision of Dr. Francesc Alías and Dr. Joan Claudi Socoró.

4. Conclusions

The GENIOVOX project has constituted a first step towards the numerical generation of diphthongs, hiatuses and fricative consonants through 3D FEM without resorting to supercomputer facilities, as well as a preliminary attempt to introduce expressiveness in vowels, specifically, modal and tense phonation types extracted from neutral and happy plus aggressive vowels.

In future works we aim at developing finite element computational strategies on dynamic 3D VTs to simulate spoken utterances containing fricative sounds as well as velar and bilabial stops. Moreover, we will keep working towards improving the naturalness of the generated expressive voice by properly modifying the glottal flow model and the VT shape to generate different voice qualities and vocal effects such as the Lombard effect or the singing formant, considering also inverse filtering techniques applied to expressive speech corpora.

5. Acknowledgments

The authors acknowledge the Agencia Estatal de Investigación (AEI) and FEDER, EU, for funding the project GENIOVOX (ref. TEC2016-81107-P).

6. References

Journal papers

[1] R. Montaño and F. Alías; "The Role of Prosody and Voice Quality in Indirect Storytelling Speech: A cross-narrator perspective in four European Languages," *Speech Communication*, vol. 88, pp. 1-16, 2017.

[2] M. Arnela and O. Guasch, "Finite element synthesis of diphthongs using tuned two-dimensional vocal tracts," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 25 (10), pp. 2013-2023, 2017.

[3] A. Pont, R. Codina, J. Baiges and O. Guasch, "Unified solver for fluid dynamics and aeroacoustics in isentropic gas flows," *Journal of Computational Physics*, 363, pp 11-29, 2018.

[4] A. Pont, O. Guasch, J. Baiges, R. Codina and A. Van Hirtum, "Computational aeroacoustics to identify sound sources in the generation of sibilant /s/," *International Journal for Numerical Methods in Biomedical Engineering*, 35 (1), e3153, pp. 1-17, 2019.
[5] S. Dabbaghchian, M. Arnela, O. Engwall and O. Guasch, "Reconstruction of vocal tract geometries from biomechanical simulations," *International Journal for Numerical Methods in Biomedical Engineering*, 35 (2), e3159, pp. 1-19, 2019.
[6] M. Freixes, M. Arnela, J.C. Socoró, F. Alías and O. Guasch, "Glottal source contribution to higher order modes in the finite element synthesis of vowels," *Applied Sciences*, 9 (21), 4535, pp. 1-12, 2019.
[7] M. Arnela, S. Dabbaghchian, O. Guasch and O. Engwall, "MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 27 (2), pp. 2173-2182, 2019.
[8] M. Freixes, F. Alías and J.C. Socoró, "A Unit Selection Text-to-Speech-and-Singing Synthesis Framework from Neutral Speech: Proof of concept," *EURASIP Journal on Audio, Speech, and Music Processing*, 2019:22, pp. 1-14, 2019.
[9] A. Pont, O. Guasch and M. Arnela, "Finite element generation of sibilants /s/ and /z/ using random distributions of Kirchhoff's vortices," *International Journal for Numerical Methods in Biomedical Engineering*, 36 (2), e3302, pp. 1-20, 2020.
[10] O. Guasch, M. Arnela and A. Pont, "Resonance tuning in vocal tract acoustics from modal perturbation analysis instead of nonlinear radiation pressure," *Journal of Sound and Vibration*, 493, 115826, 2021.
[11] S. Dabbaghchian, M. Arnela, O. Engwall and O. Guasch (2021), "Synthesis of vowels and vowel-vowel utterances using a 3D biomechanical-acoustic model," *International Journal for Numerical Methods in Biomedical Engineering, Accepted*, 2021.

Conference articles (A) and presentations (P)

[12] M. Arnela, S. Dabbaghchian, O. Guasch and O. Engwall, "A semi-polar grid strategy for the three-dimensional finite element simulation of vowel-vowel sequences", *Interspeech 2017*, August 20-24, Stockholm, (Sweden), 2017. (A)
[13] N.C. Degirmenci, J. Jansson, J. Hoffman, M. Arnela, P. Sánchez-Martín, O. Guasch and S. Ternström, "A unified simulation of vowel production that comprises phonation and the emitted sound," *Interspeech 2017*, August 20-24, Stockholm, (Sweden), 2017. (A)
[14] S. Dabbaghchian, M. Arnela, O. Engwall and O. Guasch, "Synthesis of VV utterances from muscle activation to sound with a 3D model", Interspeech *2017*, August 20-24, Stockholm, (Sweden), 2017. (A)
[15] O. Guasch and S. Ternström, "Some current challenges in unified numerical simulations of voice production: from biomechanics to the emitted sound," ISSP2017, the 11th International Seminar on Speech Production, October 16-19, Tianjin, (China), 2017. (A)
[16] R. Codina, A. Pont, J. Baiges and O. Guasch, "Split boundary conditions for computational aeroacoustics of isentropic flows," *6th European Conference on Computational Mechanics and 7th European Conference on Computational Fluid Dynamics (ECCM-ECFD 2018)*, June 11-15, Glasgow (UK), 2018. (P)
[17] M. Freixes, M. Arnela, J.C. Socoró, F. Alías and O. Guasch, "Influence of tense, modal and lax phonation on the three-dimensional finite element synthesis of vowel [A]," *IberSpeech2018*, November 21-23, Barcelona, Catalonia (Spain), 2018. (A)
[18] O. Guasch, M. Arnela and A. Pont, "Modal perturbation analysis instead of nonlinear radiation pressure to derive the area sensitivity function for resonance tuning in an axisymmetric duct with variable cross-section," *InterNoise2019*, June 16–19, Madrid, (Spain), 2019. (A)
[19] A. Pont, O. Guasch and M. Arnela, "Modal perturbation analysis instead of nonlinear radiation pressure to derive the area sensitivity function for resonance tuning in an axisymmetric duct with variable cross-section," *InterNoise2019*, June 16–19, Madrid, (Spain), 2019. (A)
[20] A. Pont, M. Arnela and O. Guasch, "Finite element generation of vowel-sibilant utterances using random distributions of Kirchhoff's vortices and simplified vocal tract geometries", *ECCOMAS YIC*, September 1-6, Krakow (Poland), 2019. (P)
[21] M. Arnela and O. Guasch, "Finite element simulation of /asa/ in a three-dimensional vocal tract using a simplified aeroacoustic source model," *ICA 2019, 23th International Congress on Acoustics*, September 9-13, Aachen (Germany), 2019. (A)
[22] O. Guasch, Survey lecture on "Realistic physics-based computational voice production," Co-contributors: Marc Arnela, Arnau Pont, Francesc Alías, Marc Freixas and Joan-Claudi Socoró, *Interspeech 2019*, September 15-19, Graz (Austria), 2019. (P)
[23] M. Freixes, M. Arnela, F. Alías, J.C. Socoró, "GlottDNN-based spectral tilt analysis of tense voice emotional styles for the expressive 3D numerical synthesis of vowel [a]," Proceedings of the 10th ISCA Speech Synthesis Workshop (SSW10), pp. 132-136, 20-22 September 2019, Vienna, Austria, 2019. (P)
[24] Marc Arnela, Oriol Guasch, Arnau Pont (2020), "Tuning MRI-based vocal tracts to modify formants in the three-dimensional finite element production of vowels", Proc. of 12th International Conference on Voice Physiology and Biomechanics (ICVPB), March Edition, Grenoble, France. (A)

Book Chapters

[25] O. Guasch, A. Pont, J. Baiges and R. Codina, "Simultaneous Finite Element Computation of Direct and Diffracted Flow Noise in Domains with Static and Moving Walls", in: Ciappi E. *et al*. (eds) Flinovia—Flow Induced Noise and Vibration Issues and Aspects-II. FLINOVIA 2017 (ISBN: 978-3-319-76779-6), pp. 179-194, Springer, Cham Hastie, 2019.

PhD Thesis

[26] M. Freixes. "Adding expressiveness to unit selection speech synthesis and to numerical voice production", PhD Thesis, La Salle – Universitat Ramon Llull, 2021.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Adverse Drug Reaction extraction on Electronic Health Records written in Spanish: A PhD Thesis overview

Sara Santiso González

IXA group, University of the Basque Country (UPV/EHU)
Manuel Lardizabal 1, 20018, Donostia
sara.santiso@ehu.eus

Abstract

The aim of this work is the automatic extraction of Adverse Drug Reactions (ADRs) in Electronic Health Records (EHRs) written in Spanish. From Natural Language Processing (NLP) perspective, this is approached as a relation extraction task in which the drug is the causative agent of a disease, the adverse reaction. This would help to increase the reporting of ADRs and their earliest possible detection, helping to improve the health of the patients.

ADR extraction from EHRs involves major challenges. First, drugs and diseases found in an EHR are often unrelated or sometimes related as treatment, but seldom as ADRs. This implies the inference of a predictive model from samples with skewed class distribution. Second, EHRs contain both standard and nonstandard abbreviations and misspellings. All this leads to a high lexical variability. Third, the Spanish count with few resources and tools to apply NLP. To cope with these challenges, we explored several ADR detection algorithms (Random Forest and Joint AB-LSTM) and representations (symbolic and dense) to characterize the ADR candidates. In addition, we assessed the tolerance of the ADR detection model to external noise such as the incorrect detection of the medical entities involved in the ADR extraction.

Index Terms: Adverse Drug Reactions, Electronic Health Records, Text mining, Supervised machine learning

1. Introduction

An ADR is defined by the World Health Organization (WHO) as 'a response to a medicine which is noxious and unintended, and which occurs at doses normally used in man' [1]. The WHO informed about the importance of reporting ADRs to understand and treat the diseases caused by drugs and, as a result, improve the patients care [2]. However, ADRs are still heavily under-reported, which makes their prevention difficult. This was the **motivation** to automatically extract ADRs on Electronic Health Records (EHRs). Given that information stored digitally by the hospitals is growing, Natural Language Processing (NLP) techniques can be used to create a system that helps the doctors to analyze the ADRs of the patients in a given EHR, facilitating the decision making process and alleviating the work-load. As a consequence, the patients' health could improve and the pharmaco-surveillance service would be informed about the detected ADRs. The ADR extraction was defined as a relation extraction task. That is, the aim is to detect ADR relations between the entities (drugs and diseases) recognized in a given text. For the ADR extraction developed in this work, we distinguished the two steps involved in this task, which were developed using a pipeline approach:

1. Medical Entity Recognition (MER) to find "drug" entities and "disease" entities. The "drug" entity encompasses either a brand name, a substance or an active ingredient and the "disease" entity encompasses either a disease, a sign or a symptom.
2. ADR detection to discover the relations between "drug" entities and "disease" entities that correspond to ADRs. The "drug" entity would be the causative agent and the "disease" entity would be the caused adverse reaction.

In the ADR extraction process, we had to overcome some **challenges** that make this supervised classification task difficult. On the one hand, the ADRs are minority relations because generally the drug and the disease are either unrelated or related as treatment and, thus, the ADRs are rare cases. This implies the inference of a predictive model from samples with skewed class distribution. On the other hand, the EHRs show multiple lexical variations. EHRs are written by experts under time pressure, employing rich medical jargon together with colloquial expressions, not always grammatical, and it is not infrequent to find misspellings and both standard and nonstandard abbreviations. In addition, ***our EHRs are written in Spanish whereas the majority of biomedical NLP research has been done in English***. The Spanish and other languages different to English count with few resources and tools to apply NLP in the medical domain. In this line, it is remarkable the recent interest in developing NLP tools for languages other than English [3]. To cope with these challenges, we explored several ADR detection algorithms, Random Forest (RF) [4] and Joint Attentive Bidirectional Long Short-Term Memory (AB-LSTM) [5], and representations to characterize the ADR candidates, symbolic and dense.

The main **objective** of this work is the creation of a model able to detect automatically ADRs in EHRs written in Spanish. This, in turn, encompasses the sub-objectives stated below:

- Detect ADRs by discovering relations between the causative drug and the caused diseases.

  The aim is to detect drug-disease pairs related as ADRs and not only the disease caused by the drug. Indicating explicitly the entities involved in an ADR can result more useful for their study.

- Discover approaches to overcome the class imbalance.

  Given that ADRs are rare events, it is frequent to find the class imbalance problem in this task. Machine learning algorithms tend to expect balanced class distributions and learning the minority class is difficult for them. For

This manuscript presents a summary of the thesis developed by Sara Santiso and supervised by Arantza Casillas and Alicia Pérez. It was defended on June 13 2019 obtaining excellent grade with Cum Laude mention.

10.21437/IberSPEECH.2021-34

this reason, our intention is to explore different techniques that could help to tackle this issue improving the ADR detection or find approaches that could be robust against imbalanced distributions of the class.

- Discover robust representations to cope with the lexical variability and the data sparsity.

  This is a challenge goal due to two factors. First, the EHRs are written during consultation time and each doctor uses different terms or expressions, producing lexical variations. Second, due to confidentiality issues, there is a lack of available EHRs. Then, our intention is to explore different representations in order to make the most of the annotated corpus.

2. Related work

One of the differentiating factors in related works dealing with ADR extraction is the definition of the task itself. In our case, it is approached as a relation extraction task between drugs and diseases (the adverse reactions). From the NLP perspective this consists in the extraction of cause-effect events between the entities. There are relevant works following this approach [6, 7, 8, 9]. Alternatively, some authors considered ADR extraction as the identification of the caused disease (the adverse reaction), that is, a sub-class of MER [10, 11, 12, 13] or refer to ADR extraction as the detection of presence (or absence) of ADRs in a document [14, 15, 16].

ADR extraction was applied to several textual genres such as social media, scientific publications or EHRs. Among the works developed with EHRs we distinguished three approaches. The earliest attempts made use of symbolic features (e.g. word-forms, lemmas and POS tags) to represent the ADRs and employed traditional classifiers [17, 18]. Next, with the growing distributional semantics, word-embeddings were introduced and main trends in characterizations turned from symbolic into dense spaces [19]. Finally, deep neural networks re-emerged as state-of-the-art classification approaches [20, 21].

For Spanish, we found the SpanishADRCorpus [10] labeled with drugs and effects as entities and drug indications and adverse drug reactions as relations. It is composed by 400 documents gathered from ForumClinic, a health network website in Spanish. Note that this corpus was employed for ADR extraction with techniques based on rules and unsupervised methods [22, 23, 24].

3. Methods

3.1. Corpora

In this work we employed three annotated corpora that involve EHRs written in Spanish from two hospitals within Osakidetza. The gold standard corpus (IxaMed-GS) contains 75 EHRs from the Galdakao hospital [25]. The cross hospital corpus (IxaMed-CH) contains 267 EHRs from the Galdakao and Basurto hospitals. It contains the EHRs of IxaMed-GS. The extended corpus (IxaMed-E) contains 463 EHRs from the Galdakao and Basurto hospitals. It contains some EHRs of IxaMed-CH but not of IxaMed-GS. In order to infer and evaluate the models, each corpus was divided in train, development and test sets randomly selected without replacement. Parameters were tuned by training with the train set and evaluating on the dev set (train vs dev). With those parameters the model was trained with the union of the train and dev sets and evaluated on the test set (train∪dev vs test). While the positive ADR relations were those manually annotated by the experts, the negative relations were created by combining all the Disease group and Allergy entities with all the Drug group entities present in each document. Table 1 shows the quantitative description of these corpora: number of documents, word-forms, vocabulary, Out-Of-Vocabulary (OOV) words and medical entities that the experts manually tagged together with the number of ADR relations of each class. Note that the OOVs are words of the evaluation set that were not seen in the training set. The OOV of the Dev set are calculated with respect to the vocabulary of the Train set and the OOV of the Test set are calculated with respect to the vocabulary of the Train and Dev sets.

Table 1: *Quantitative description of the corpora. Positive relations (⊕) refer to ADRs while negative relations (⊖) refer to non-ADRs.*

Corpus			Partition		
			Train	**Dev**	**Test**
IxaMed-GS	**Documents**		41	17	17
	Word-forms		20,689	11,246	9,698
	Vocabulary		4,934	–	–
	OOV		–	1,526	979
	Entities	**Drug**	503	346	354
		Disease	1,341	737	629
	Relations	⊕	53	30	27
		⊖	231	134	173
IxaMed-CH	**Documents**		157	55	55
	Word-forms		91,088	34,004	33,171
	Vocabulary		13,809	–	–
	OOV		–	2,628	2,280
	Entities	**Drug**	2,436	943	887
		Disease	6,828	2,328	2,473
	Relations	⊕	197	79	62
		⊖	2,162	366	559
IxaMed-E	**Documents**		279	92	92
	Word-forms		138,695	47,487	43,858
	Vocabulary		18,003	–	–
	OOV		–	3,182	2,735
	Entities	**Drug**	3,474	1,128	1,122
		Disease	10,894	3,831	3,387
	Relations	⊕	332	113	82
		⊖	12,877	5,312	3,756

3.2. ADR detection approaches

Firstly, we used three approaches for ADR extraction, with alternative characterizations and classification algorithms, using the IxaMed-GS corpus.

- ADR detection with symbolic representations and RF

  We explored symbolic characterizations with the traditional classifier RF [26, 27, 28, 29, 30]. First, we tackled the detection of intra-sentence as well as inter-sentence ADRs. In order to overcome the class imbalance we tried different techniques: sampling, cost-sensitive learning, ensemble learning and one-class classification. Finally, we restricted the task to the same sentence to reduce the level of imbalance. Note that we also explored two approaches to detect negated entities automatically [31, 32]. These would be used to discard negative ADR candidates.

- ADR detection with dense representations and RF

  We explored dense characterizations created from embeddings that were used together with the RF classifier overcoming the class imbalance [33]. Instead of using an embedding for each word, we built a single context-aware embedding (dense representation created taking into account the embeddings of the context-words). Furthermore, we proposed different smoothing techniques that were applied to the dense representations to improve the proximity between semantically related words. These techniques are direction cosines, truncation, Principal Component Analysis (PCA) and clustering.

- ADR detection with dense representations and Joint AB-LSTM

  We used dense characterization automatically inferred by the Joint AB-LSTM network [34]. Specifically, in the Joint AB-LSTM two Bi-LSTM are trained. We compared different pooling strategies such as max, average and attention pooling separately and also their combinations. We explored the use of word-forms and lemmas as core-features. Furthermore, we explored the techniques to overcome the class imbalance suited for neural networks (re-sample, re-sample per batch and cost-sensitive learning).

3.3. Tolerance of ADR detection to noise

Finally, with the best performing approach, we used different corpora (IxaMed-GS, IxaMed-CH and IxaMed-E) with a higher number of examples and slightly variations in the sub-domains (EHRs from different hospitals with different services or specializations). In addition, we also analyzed the influence in the ADR extraction of a real system for the automatic detection of medical entities. Specifically, we employed Conditional Random Fields (CRF) [35] as classifier.

4. Results

4.1. Results for the ADR detection approaches

Developing the ADR detection with symbolic representations and RF, the best results were obtained at sentence level with the application of re-sample. Restricting the ADR detection to sentence level alleviated drastically the class imbalance problem, reducing the imbalance ratio from 1:222 to 1:4. Among all the hand-crafted features used in our symbolic representation, the 20 most relevant features for the intra-sentence scope were the word-forms and lemmas of the entities and their contexts. By contrast, the distances are the most relevant ones when inter- and intra-sentence scope is considered.

With dense representations and RF the best performing model was created with the application of director cosines, truncation and PCA. The embeddings were generated with a corpus of EHRs using GloVe. We compared the results obtained using the concatenation of words with those obtained replacing the words by their corresponding embeddings and it led to significant improvements. We also observed that the smoothing techniques outperformed their corresponding non-smoothed counterpart.

In the case of dense representations and Joint AB-LSTM, the best results were obtained without tackling the class imbalance, using a lemmatized version of the embeddings, Batch Normalization and combining max pooling and attentive pooling. We used a Feed Forward Neural Network (FFNN) as baseline, a simplified version of the Joint AB-LSTM that skips the Bi-LSTM layer. This was outperformed by the Joint AB-LSTM. We found that Batch Normalization was helpful and lemmatization was effective. According to the results, it seemed as if max and attention pooling complimented each other.

Table 2 shows these results, which demonstrate that the dense representation resulted useful since the f-measure of the positive class obtained with the symbolic representation improved in all the cases. Furthermore, the abstract representation automatically inferred by the Joint AB-LSTM was even better. For more detailed see [30, 34, 34].

Table 2: *Results of each best performing approach (symbolic + RF, dense + RF, dense + Joint AB-LSTM) for the IxaMed-GS corpus.*

		train vs dev			train∪dev vs test			Class
		P	**R**	**F**	**P**	**R**	**F**	
symbolic	RF	54.5	40.0	46.2	34.0	59.3	43.2	⊕
		87.3	92.5	89.9	92.8	82.1	87.1	⊖
		81.3	82.9	81.9	84.9	79.0	81.2	W. Avg.
		82.9	82.9	82.9	79.0	79.0	79.0	Micro Avg.
		70.9	66.3	68.0	63.4	70.7	65.2	Micro Avg.
dense	RF	54.8	76.7	63.9	47.4	66.7	55.4	⊕
		94.3	85.8	89.8	94.4	88.4	91.3	⊖
		87.0	84.1	85.1	88.1	85.5	86.5	W. Avg.
		84.1	84.1	84.1	85.5	85.5	85.5	Micro Avg.
		74.5	81.2	76.9	70.9	77.6	73.4	Micro Avg.
	Joint AB-LSTM	87.2	67.8	**76.3**	72.4	71.4	**71.9**	⊕
		93.2	97.8	95.4	95.3	95.5	95.4	⊖
		92.1	92.3	91.9	92.0	92.1	92.0	W. Avg.
		92.3	92.3	92.3	92.1	91.8	92.7	Micro Avg.
		90.2	82.8	85.8	83.8	83.4	83.6	Micro Avg.

4.2. Results for the tolerance of ADR detection to noise

Analyzing the tolerance of ADR detection to noise we observed that despite of increasing the class imbalance and the sub-domains, the results improved as the size of the corpus increased. This is shown in Table 3, where the best results were obtained with IxaMed-E, the largest corpus.

In addition, Table 4 gives the results obtained with CRF as entity recognizer. The f-measure of the positive class suggests that ADR relations are scarce and missing some entities does not have an impact in the results while they are important in clinical practice.

5. Discussion

In this work we made a step ahead in the development of NLP methods that deal with ADR extraction defined as relation extraction task between a causative drug and the adverse reaction. We observed that the combination of approaches to tackle the high class imbalance, precisely sampling and cost-sensitive learning, was beneficial in the context of inter- and intra-sentence ADR extraction. We also observed that class imbalance can be, somehow, tackled in intra-sentence ADR extraction.

We experimentally corroborated that to deal with lexical variability, context-aware embeddings are useful to preserve the lexical nuances in this domain. Furthermore, to alleviate the influence that the lack of training samples might have in the

Table 3: *Results of the best performing approach (dense + Joint AB-LSTM) with each corpus (IxaMed-GS, IxaMed-CH, IxaMed-E).*

	train vs dev			train∪dev vs test			Class
	P	**R**	**F**	**P**	**R**	**F**	
IxaMed-GS	87.2	67.8	76.3	72.4	71.4	71.9	⊕
	93.2	97.8	95.4	95.3	95.5	95.4	⊖
	92.1	92.3	91.9	92.0	92.1	92.0	W. Avg.
	92.3	92.3	92.3	92.1	91.8	92.7	Micro Avg.
	90.2	82.8	85.8	83.8	83.4	83.6	Micro Avg.
IxaMed-CH	89.3	69.2	77.9	76.0	70.9	73.3	⊕
	94.1	98.3	96.2	96.1	96.9	96.5	⊖
	93.2	93.4	93.1	93.7	93.8	93.7	W. Avg.
	93.4	93.4	93.4	93.8	93.8	93.8	Micro Avg.
	91.6	83.7	87.0	86.1	83.9	84.9	Micro Avg.
IxaMed-E	90.3	71.8	**79.9**	74.4	76.0	**75.2**	⊕
	94.7	98.5	96.6	96.5	96.2	96.3	⊖
	94.0	94.2	93.9	93.7	93.6	93.7	W. Avg.
	94.2	94.2	94.2	93.6	93.6	93.6	Micro Avg.
	92.6	85.2	88.3	85.4	86.1	85.8	Micro Avg.

Table 4: *Results of the best performing approach (dense + Joint AB-LSTM) with the IxaMed-E corpus, evaluated using the automatic entities.*

train vs dev			train∪dev vs test			Class
P	**R**	**F**	**P**	**R**	**F**	
96.4	60.7	74.5	86.2	53.1	65.7	⊕
92.9	99.5	96.1	93.6	98.8	96.1	⊖
93.5	93.3	92.6	92.6	93.0	92.3	W. Avg.
93.3	93.3	93.3	93.0	93.0	93.0	Micro Avg.
94.7	80.1	85.3	89.9	75.9	80.9	Micro Avg.

quality of the inferred dense representations, we proposed the use of smoothing techniques. We observed that dense spaces of lemmas also helped to tackle the lexical variability. In fact, lemmatization was particularly effective in the neural networks used for ADR extraction.

In addition, we corroborated that the Joint AB-LSTM is able to cope with these types of noise although, naturally, there is a small decrease in its performance due to the missed entities involved in the ADR pairs.

The main **contribution** of this work is that the ADR extraction was developed using EHRs written in Spanish. To the best of our knowledge, for ADR extraction in texts written in Spanish, we are the first employing EHRs. In related works we observed that there is a corpus written in Spanish, SpanishADR-Corpus [10], that was employed to develop several works for ADR detection. However, this corpus is not composed by EHRs and the authors employed techniques based on rules and unsupervised methods instead of supervised learning.

6. Conclusions and future work

ADRs are rare events, then, supervised classification algorithms tend to be biased and learning to predict the minority class is complex. The application of approaches to overcome the **class imbalance** improves the performance of the ADR detection model to find **inter-** as well as **intra-sentence** ADRs. However, the results are considerably better in the intra-sentence scope than in the inter-sentence scope.

A key issue in the extraction of ADRs is the operative **characterization** of events. With regard to initial symbolic characterizations, if both inter-sentence and intra-sentence relations are taken into account, features related to the distances between the entities involved result relevant for the task. If the ADR detection is focused on intra-sentence ADRs, the word-forms and the lemmas of the entities and their contexts are more relevant. NLP rapidly evolved towards dense characterizations. Dense representations have the strength of exploiting semantic relatedness in dense low dimensional spaces. This is an important factor in our task to cope with lexical variability. We corroborated that dense representations outperform symbolic ones and it seemed as if the model gains generalization ability.

Another important factor is the **classification approach**. In this work we compared a traditional supervised classification approach (RF) and an emerging technique based on deep neural networks (Joint AB-LSTM) and found that Joint AB-LSTM outperformed RF. We speculated about the reasons behind. An outstanding difference between traditional and neural approaches rests on the generation of the inherent characterization for the instances. While traditional approaches make use of hand-crafted features (either in their symbolic or embedded as dense representations), neural approaches infer, automatically, abstract features. Nevertheless, we found that FFNN did not outperform the RF when the instances were characterized with smoothed embeddings. Our hypothesis to explain that Joint AB-LSTM outperform RF is that the information captured from the context is crucial in relation extraction. While RF exploits the context in a static way, Joint AB-LSTM can leverage the context dynamically. Furthermore, we observed, empirically, that Joint AB-LSTM networks are less sensitive to class imbalance than RF.

Variations in the size and domain of the corpus have an **effect in the performance** of the ADR detection model. To be precise, the larger the corpus the better the results. Regarding the variations associated to different sub-domains introduced by the use of EHRs of different hospitals, Joint AB-LSTM resulted robust. Needless to say, the errors propagated from the MER step affect the ADR detection. Missing entities lead to undiscovered relations. However, the drop in performance is not as dramatic as we expected.

ADR extraction is a major issue in pharmaco-surveillance and documentation. So far, the systems tend to focus on the detection of drug-disease pairs located in the same sentence (intra-sentence relations). Nevertheless, EHRs have implicit information that might reveal underlying relations (e.g. information in the antecedents might be relevant to guess the causes for an adverse event). That is, as **future work** an effort should be made to detect inter-sentence relations both, explicitly and implicitly stated. Moreover, we can work in developing entity recognition and relation extraction simultaneously, using a joint model to avoid the propagation of pipeline errors.

7. Acknowledgements

The authors would like to thank the staff of the Pharmacy and Pharmacovigilance services of the Galdakao-Usansolo and Basurto hospitals. This work was partially funded by the Spanish Ministry of Science and Innovation (PROSAMED: TIN2016-77820-C3-1-R) and the Basque Government (BERBAOLA: KK-2017/00043, Predoctoral Grant: PRE 2018 2 0265).

8. References

[1] World Health Organization, "Safety of medicines: a guide to detecting and reporting adverse drug reactions: why health professionals need to take action," pp. 1–16, 2002.

[2] ——, "The importance of pharmacovigilance," pp. 1–52, 2002.

[3] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, "Clinical natural language processing in languages other than English: opportunities and challenges," *Journal of Biomedical Semantics*, vol. 9, no. 1, pp. 1–13, 2018.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] S. K. Sahu and A. Anand, "Drug-drug interaction extraction from biomedical texts using long short-term memory network," *Journal of Biomedical Informatics*, vol. 86, pp. 15–24, 2018.

[6] H. Gurulingappa, A. Mateen-Rajpu, and L. Toldo, "Extraction of potential adverse drug events from medical case reports," *Journal of Biomedical Semantics*, vol. 3, no. 1, pp. 1–10, 2012.

[7] F. Li, D. Ji, X. Wei, and T. Qian, "A transition-based model for jointly extracting drugs, diseases and adverse drug events," in *2015 IEEE International Conference on Bioinformatics and Biomedicine*, 2015, pp. 599–602.

[8] J. Legrand, Y. Toussaint, C. Raïssi, and A. Coulet, "Syntax-based transfer learning for the task of biomedical relation extraction," in *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, 2018, pp. 149–159.

[9] B. He, Y. Guan, and R. Dai, "Classifying medical relations in clinical text via convolutional neural networks," *Artificial Intelligence in Medicine*, vol. 93, pp. 43–49, 2019.

[10] I. Segura-Bedmar, R. Revert, and P. Martínez, "Detecting drugs and adverse events from Spanish social media streams," in *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, April 2014, pp. 106–115.

[11] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez, "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," *Journal of the American Medical Informatics Association*, vol. 22, no. 3, pp. 671–681, 2015.

[12] G. Stanovsky, D. Gruhl, and P. Mendes, "Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, vol. 1, April 2017, pp. 142–151.

[13] S. Gupta, S. Pawar, N. Ramrakhiyani, G. K. Palshikar, and V. Varma, "Semi-supervised recurrent neural network for adverse drug reaction mention extraction," *BMC Bioinformatics*, vol. 19, no. 8, pp. 1–7, 2018.

[14] I. Karlsson, J. Zhao, L. Asker, and H. Boström, "Predicting adverse drug events by analyzing electronic patient records," in *Conference on Artificial Intelligence in Medicine in Europe*, 2013, pp. 125–129.

[15] J. Zhao, A. Henriksson, L. Asker, and H. Boström, "Predictive modeling of structured electronic health records for adverse drug event detection," *BMC Medical Informatics and Decision Making*, vol. 15, no. 4, pp. 1–15, 2015.

[16] S. Friedrich and H. Dalianis, "Adverse drug event classification of health records using dictionary based pre-processing and machine learning," in *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, 2015, pp. 121–130.

[17] E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, K. Waki, and K. Ohe, "Extraction of adverse drug effects from clinical records." in *MedInfo*, 2010, pp. 739–743.

[18] Y. Miura, E. Aramaki, T. Ohkuma, M. Tonoike, D. Sugihara, H. Masuichi, and K. Ohe, "Adverse-effect relations extraction from massive clinical records," in *Proceedings of the Second Workshop on NLP Challenges in the Information Explosion Era*, 2010, pp. 75–83.

[19] A. Henriksson, M. Kvist, H. Dalianis, and M. Duneld, "Identifying adverse drug event information in clinical notes with distributional semantic representations of context," *Journal of Biomedical Informatics*, vol. 57, pp. 333–349, 2015.

[20] Y. Luo, "Recurrent neural networks for classifying relations in clinical notes," *Journal of Biomedical Informatics*, vol. 72, pp. 85–95, 2017.

[21] D. Raj, S. Sahu, and A. Anand, "Learning local and global contexts using a convolutional recurrent network model for relation classification in biomedical text," in *Proceedings of the 21st Conference on Computational Natural Language Learning*, 2017, pp. 311–321.

[22] S. de la Peña, I. Segura-Bedmar, P. Martínez, and J. L. Martínez, "ADRSpanishTool: a tool for extracting adverse drug reactions and indications," *Procesamiento del Lenguaje Natural*, vol. 53, pp. 177–180, 2014.

[23] I. Segura-Bedmar, S. de la Peña, and P. Martınez, "Extracting drug indications and adverse drug reactions from Spanish health social media," in *Proceedings of BioNLP*, 2014, pp. 98–106.

[24] I. Segura-Bedmar, P. Martínez, R. Revert, and J. Moreno-Schneider, "Exploring Spanish health social media for detecting drug effects," *BMC Medical Informatics and Decision Making*, vol. 15, no. 2, pp. 1–9, 2015.

[25] M. Oronoz, K. Gojenola, A. Pérez, A. Díaz de Ilarraza, and A. Casillas, "On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions," *Journal of Biomedical Informatics*, vol. 56, pp. 318–332, 2015.

[26] S. Santiso, A. Casillas, A. Perez, M. Oronoz, and K. Gojenola, "Adverse drug event prediction combining shallow analysis and machine learning," in *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, 2014, pp. 85–89.

[27] S. Santiso, A. Casillas, A. Pérez, M. Oronoz, and K. Gojenola, "Document-level adverse drug reaction event extraction on electronic health records in Spanish," *Procesamiento del Lenguaje Natural*, no. 56, pp. 49–56, 2016.

[28] A. Casillas, A. Pérez, M. Oronoz, K. Gojenola, and S. Santiso, "Learning to extract adverse drug reaction events from electronic health records in Spanish," *Expert Systems with Applications*, vol. 61, pp. 235–245, 2016.

[29] A. Casillas, A. D. de Ilarraza, K. Fernandez, K. Gojenola, M. Oronoz, A. Pérez, and S. Santiso, "IXAmed-IE: On-line medical entity identification and ADR event extraction in Spanish," in *2016 IEEE International Conference on Bioinformatics and Biomedicine*, 2016, pp. 846–849.

[30] S. Santiso, A. Casillas, and A. Pérez, "The class imbalance problem detecting adverse drug reactions in electronic health records," *Health Informatics Journal*, pp. 1–11, 2018.

[31] S. Santiso, A. Casillas, A. Pérez, and M. Oronoz, "Medical entity recognition and negation extraction: Assessment of NegEx on health records in Spanish," in *International Work-Conference on Bioinformatics and Biomedical Engineering*, 2017, pp. 177–188.

[32] ——, "Word embeddings for negation detection in health records written in Spanish," *Soft Computing*, pp. 1–7, 2018.

[33] S. Santiso, A. Pérez, and A. Casillas, "Smoothing dense spaces for improved relation extraction between drugs and adverse reactions," *International journal of medical informatics*, vol. 128, pp. 39–45, 2019.

[34] ——, "Exploring joint ab-lstm with embedded lemmas for adverse drug reaction discovery," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–8, 2018.

[35] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, vol. 1, 2001, pp. 282–289.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Design and Evaluation of Mobile Computer-Assisted Pronunciation Training Tools for Second Language Learning: a Ph.D. Thesis Overview

Cristian Tejedor-García[1], *Valentín Cardeñoso-Payo*[1], *David Escudero-Mancebo*[1]

[1]ECA-SIMM Research Group, Department of Computer Science, University of Valladolid, Spain

cristian@infor.uva.es

Abstract

Recent advances on speech technologies (automatic speech recognition, ASR, and text-to-speech, TTS, synthesis) have led to their integration in computer-assisted pronunciation training (CAPT) tools. However, pronunciation is an area of teaching that has not been developed enough since there is scarce empirical evidence assessing the effectiveness of CAPT tools and games that include ASR/TTS. In this manuscript, we summarize the findings presented in Cristian Tejedor-García's Ph.D. Thesis (University of Valladolid, 2020). In particular, this dissertation addresses the design and validation of an innovative CAPT system for smart devices for training second language (L2) pronunciation at the segmental level with a specific set of methodological choices, such as the inclusion of ASR/TTS technologies with minimal pairs, learner's native–foreign language connection, a training cycle of exposure–perception–production, and individual/social approaches. The experimental research conducted applying these methodological choices with real users validates the efficiency of the CAPT prototypes developed for the four main experiments of this dissertation about English and Spanish as L2. We were able to accurately measure the relative pronunciation improvement of the individuals who trained with them. Expert raters on phonetics' subjective scores and CAPT's objective scores showed a strong correlation, being useful in the future to be able to assess a large amount of data and reducing human costs.

Index Terms: Computer-assisted pronunciation training (CAPT), second language (L2) pronunciation, automatic speech recognition (ASR), text-to-speech (TTS), autonomous learning, automatic assessment tools, learning environments, mobile learning game, minimal pairs

1. Introduction

Computer-assisted pronunciation training (CAPT) tools are ideal and attractive for language learning since they provide learner autonomy and individualized and tireless instruction [1]. When these tools are well designed, it is expected to provide learners with more pronunciation practice than in traditional courses [2], real time and consistent feedback [3], and the automatic measurement of pronunciation quality [4].

In this manuscript, we present a summary of the main findings of the Ph.D. Thesis defended by Cristian Tejedor-García on the 30th of September, 2020 at University of Valladolid (Spain), *cum laude* and internal mentions. We designed and evaluated different CAPT experiments in which the main second languages (L2) were English and Spanish. We also shed light about how far we are with respect to the ideal CAPT scenario presented in the previous paragraph. Interested readers can find the numerical and statistical results of the experimentation in the complete dissertation manuscript [5].

This work was supported in part by the Ministerio de Economía y Empresa (MINECO), in part by the European Regional Development Fund FEDER under Grant TIN2014-59852-R and Grant TIN2017-88858-C2-1-R, in part by the Consejería de Educación de la Junta de Castilla y León under Grant VA050G18 and Grant VA145U14, and in part by the University of Valladolid (Ph.D. Research Grant 2015 and MOVILIDAD DOCTORANDOS UVa 2019).

1.1. Motivation

There are approximately 2.7 thousand million L2 speakers worldwide [6] and e-learning is an attractive and emerging option as a complement for one-to-one tutoring and classroom courses [7, 8]. Mobile learning applications and games are appearing for language learning for two reasons. First, they offer anytime, anywhere, and high intensity training [9]; and the quality of automatic speech recognition (ASR) [10] and text-to-speech (TTS) synthesis modules [11] that they include has considerably been improved in the last decade. However, there have been few attempts to empirically measure the effectiveness of mobile CAPT tools with ASR and TTS systems [2].

The reasons why it is not easy to measure the effectiveness of CAPT tools are mainly five. First, the similarities and differences of the native language (L1) and L2 must be taken into account; the training activities should be adapted to each learner; and the feedback provided could be insufficient or too difficult to understand by the learners [12]. Also, the automatic measurement of the quality of learner's pronunciation is not trivial and the ASR/TTS modules selected must be tested previously. Fourth, game elements and interaction with other learners might deviate from the goal of pronunciation improvement or discourage learners [13]. Finally, although multidisciplinarity is in most cases an advantage, CAPT needs to deal with some issues of communication between experts from different areas since they focus on different aspects when evaluating [1]. Thus, the current challenge is to carefully design and adapt an effective CAPT system with not only such speech technologies, but also with a training methodology, a pronunciation improvement assessment, and a corrective feedback strategy, according to learner's L1 and L2.

1.2. Research Objectives and Questions

The main objective of this thesis is to design and evaluate a CAPT tool for smart devices which incorporates current ASR and TTS technologies; helping learners to work autonomously, at their own pace, and with the possibility of providing real-time feedback. It is divided into four specific research objectives:

RO1. To analyze and define a set of activities, protocols, and motivational elements for the improvement of L2 pronunciation with a CAPT system which integrates ASR and TTS technologies.

RO2. To select the most appropriate metrics for the assessment of the learner's pronunciation level.

10.21437/IberSPEECH.2021-35

RO3. To design a semi-automatic method supervised by experts for obtaining a specific set of minimal pairs adapted to L2 pronunciation problems, according to the learner's L1 and to the limitations of the ASR and TTS technologies.

RO4. To select and design a CAPT system with current ASR and TTS technologies that provides an individualized feedback to the learner for improving L2 pronunciation.

In order to carry out the experimental procedure, three research questions were identified to validate the research objectives, categorized by topics. The first topic was related to the feasibility of current speech technology (ASR and TTS systems) integration in CAPT tools:

RQ1. Can current ASR and TTS systems be successfully used in a non-obstructive way in the CAPT tool developed?

Issue 1.1. Can current ASR and TTS systems help to assess different groups of learners according to their L2 pronunciation level in the CAPT tool developed?

The second topic referred to the implications of the training methodology with CAPT tools in learner's pronunciation improvement:

RQ2. To what extent can methodologically sensitive design issues, such as the use of exercises based on minimal pairs within the training activities cycle proposed in the CAPT tool developed affect user's pronunciation improvement?

Issue 2.1. Can a relative improvement in the learner's pronunciation be assessed after using the CAPT tool?

Issue 2.2. If any, is there a relevant pronunciation improvement from a quantitative point of view?

Issue 2.3. Does the tool reveal what the real difficulties of the users are (most difficult sounds/training activities)?

Finally, the last research question aimed at answering how game elements and social approaches affect learner's implication in pronunciation training with CAPT tools:

RQ3. To what extent can gamified versions of the tool affect user's motivation, performance, and learning?

1.3. Research Methodology

An experimental research [14] was conducted for the whole experimentation process to accomplish the objectives and give answers to the research questions proposed in this thesis, with a multidisciplinary group of researchers and experts.

Five phases were followed in each experimental iteration. First, the research problem was identified. The process started by clearly identifying the problems that will be addressed during the research process, examining the existing solutions in the state-of-the-art, and considering what possible methods will affect such solutions. Second, it was carefully planned and devised the experimental research study to test the research objectives and questions: (1) the participants were selected, that is, target population, enrollment rules, sample size, and groups; (2) different metrics were defined to measure the research variables from the data results gathered from the instruments; (3) the assessment protocol was defined, the research variables were measured before, during, and after performing the training activities; and (4) a CAPT tool was developed for each experiment of the thesis. Third, the experiment was conducted. At the beginning, the participants' groups were established. Then, each user performed the activities defined for her/his group in the previous phase, and the experimental data related to the variables of the study was collected with specific instruments for each experiment. Fourth, the data gathered was analyzed indicating the relevant indicators in order to corroborate whether the experiment was successful. Finally, the most relevant results were shared and published in scientific journals and conferences by means of articles, abstracts, show and tell demonstrations, and presentations.

2. Experimental Procedure

Given the complexity of the task, it was prudent to address it in an incremental and evolutive approach to facilitate that the results of each one of the experiments could be analyzed to change and refine the design of the next one (see Figure 1).

The evolution of the prototypes was aligned with two main focuses for the different methodologies followed to address the L2 pronunciation training through four experiments and six prototypes[1]. The first focus, *Game* (at the bottom of the figure), proposed incorporating gamification elements and social strategies; whereas the second focus, *Guided*, followed an individual approach with guided training instructions.

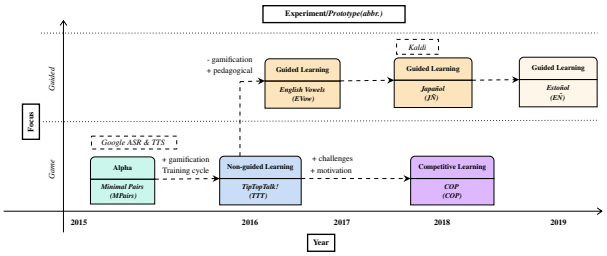


Figure 1: *Evolution diagram of the experiments and prototypes.*

2.1. Game-based Experiments

The first experiment was called **Alpha**, and the prototype developed was **Minimal Pairs**. It was a proof of concept, a starting point for checking the viability of including state-of-the-art general–purpose ASR and TTS technologies (Google) in an L2 pronunciation training protocol with minimal pairs (two similar sounding words that differ in only one phonological element and have distinct meanings); and their possible weaknesses and limitations. We also wanted to find reliable metrics to assess pronunciation tasks (see an example in Figure 2); and gather opinions from the users of a focus group session for future experiments

The number of uses of the speech technologies, success rates, and the position of the target word in the ASR hypotheses list (RO1), proved to be valid metrics (RO2) to relate the performance results with minimal pairs activities (RO3) to the proficiency level declared of each one of the participants of the three different groups (RQ1): natives, non-natives advanced-level learners, and non-natives beginner-level ones. The use of TTS by non-natives was isolated and voluntary; though no significant improvements in success rates were obtained after its use. Besides, some limitations in ASR technology were found, as native learners did not successfully complete all production activities. Finally, the opinions in the focus group session about the tool suggested new activities, feedback techniques, and motivational elements to consider for future experiments.

After checking the viability of the application of commercial ASR/TTS speech technologies and minimal pairs for a L2 pronunciation training protocol in the first experiment, we

[1]The ongoing Estoñol prototype started at the end of this Ph.D. Thesis and is still active with the collaboration of the University of Tartu, Estonia.

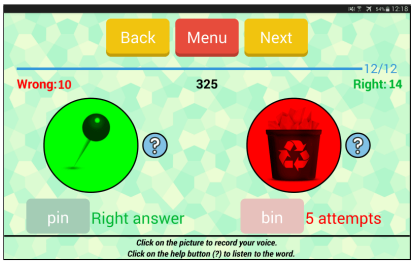


Figure 2: *Minimal Pairs prototype CAPT tool's interface [15].*

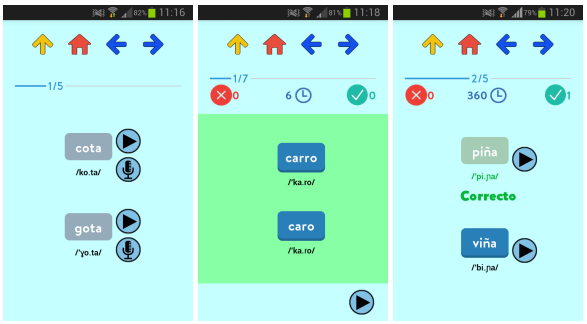


Figure 3: *TipTopTalk! prototype CAPT tool's interfaces [16].*

wanted to assess the possible pronunciation improvement after training for a longer period of time. Following the recommendations of the users, we included more training activities in the second experiment, **Non-guided Learning**: exposure, discrimination, and production tasks. We also included more gamification elements to motivate players to keep on playing. so we designed an individual competition **TipTopTalk!** in which learners played with the tool (see Figure 3), installed in their own smart devices.

The activities of perception (with TTS) were the most trained ones (RO1, RQ1), which led to a general improvement in the ability of the players (RO2, RO3, RO4, RQ2). They tended to train the easiest activities for achieving positive results. However, despite the introduction of gamification elements (RQ3) and the improvement in discrimination skills, a production improvement stagnation and a loss of interest was detected in the most proficient players. Finally, the answers to questionnaires revealed good usability results and the requirement of new feedback techniques in future work.

The training approach in the previous experiments was individual in all cases. Participants performed different activities with the tool without sharing tasks. At this point, we considered the possibility of including not only game elements, but also challenges between participants to improve their motivation and performance, improving the results of participation registered in the prototype, TipTopTalk! In order to do so, we designed a second version of the game, called **COP**, with common game rules in which each player participated in challenges with others via a mobile application (see Figure 4). In these challenges with ASR, TTS and minimal pairs, the players achieved points, and climbed up a leaderboard, while we gathered a large amount of speech and behavior data. We wanted to measure the effects of an intensive use of the game (high participation in challenges) on the most active user's motivation, performance, and learning improvement.

The lessons learned of this last game-based prototype, COP,

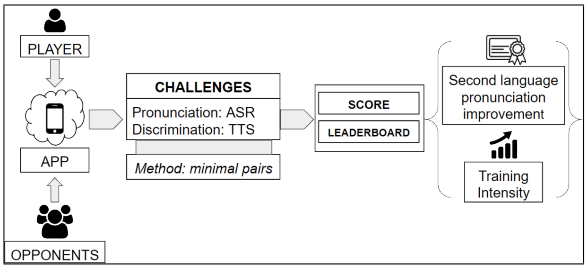


Figure 4: *COP prototype's conceptual view [17].*

were three. First, results showed an intensive practice (RQ1, RO1, RO2, RO3, RO4) supported by a significant quantity of data gathered, activities, and game days (RQ3). The most active and motivated players achieved significant pronunciation improvement results (RQ2). These outcomes were supported by the answers to the questionnaires and focus groups.

2.2. Guided-based Experiments

The second focus of the experimentation started with the **Guided Learning** experiment and **English Vowels** prototype. Both the stagnation detected in the second prototype TipTopTalk! on the most proficient learners in production improvement and their interest on performing the easiest activities, motivated this new focus of research, more pedagogical, guided, and individualized. The main objective was to train users' pronunciation by guiding them through a CAPT tool with personalized and more accurate feedback, in a controlled training protocol based on user's results and with a pre/post test strategy with minimal pairs. We included an in-classroom group to measure the proficiency improvement without a tool but with the same training activities with a teacher. Finally, we also wanted to measure the correlation between human ratings and ASR ones.

Then, we wanted to take a step further the English Vowels prototype, so we looked for alternatives to the general–purpose Google ASR system to assess a greater quantity of utterances, developing an in–house ASR system with Kaldi to address the question of how these general and specific–purpose ASR systems can deal with the assessment of minimal pairs in the field of CAPT. As a consequence, we designed **Japañol**, a CAPT tool for training Spanish pronunciation for Japanese native learners. Figure 5 shows the main training steps in the Japañol tool.

Results of both prototypes showed an intensive and effective practice (RQ1) that led to a significant improvement in pronunciation skills (RQ2). This improvement was higher in the CAPT tool group (RO1, RO3). In addition, the feedback offered in the training turned out to be effective (RO4), since those who followed it obtained better results than those who did not. Also we found strong correlations between the scores of the CAPT tool, human raters' scores, and ASR ones in the post-test (RO2). Finally, the tailored Kaldi–based ASR developed for the Japañol prototype showed that it can be as useful as a general–purpose ASR for assessing minimal pairs in CAPT tools (RO4) opening the possibility of designing a detailed study of pronunciation errors and results for future projects.

3. Conclusions

The main contributions of this Ph.D. Thesis are the following. First, we concluded that ASR and TTS technologies were ef-

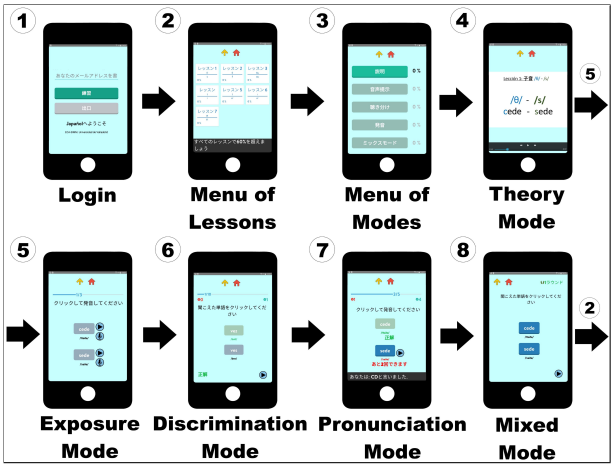


Figure 5: *Standard flow to complete a lesson in Japañol [18].*

fectively incorporated in a non-obstructive way to the L2 CAPT tools developed. Also, they proved to be a very useful didactic instrument that can be used complementary with other forms of second language acquisition.

Regarding training methodology, the versions of the tool designed and validated, and the methodological choices that they implemented, allowed us to measure the pronunciation improvement of the individuals who trained with them. Also, different corrective feedback techniques proved to be useful to overcome the proposed training activities. Our automatic data gathering method allowed us to provide specific feedback to users instantly and to analyze all results at the end of the experiments. Moreover, the minimal pair lists were elaborated with a novel semi-automatic protocol, taking into account learner's L1 and L2 and the specific ASR and TTS technologies [19].

We reported on positive results from both subjective and objective techniques for assessing pronunciation improvement during and after using the tools. The CAPT values highly correlated ($r > 0.8$) to expert evaluation. Results reported from the guided-based prototypes English Vowels and Japañol were very promising since the students achieved significant pronunciation improvement values (around one point on a ten-point grading scale), being higher in those who worked with the tool. In the COP prototype, the most active learners achieved the best pronunciation improvement and motivation results.

Finally, the game elements included in the tools proved capable of motivating learners to keep on playing on their own. The COP challenges proved to be a positive motivational factor since the most active learners made an intensive use of the game and significantly improved their L2 pronunciation along time.

As main limitations of this work, we assume that a closer comparative research between the activity and results of the tools and those of human-led instruction, such as video-taping or log monitoring might have helped us to obtain a more detailed knowledge on the possibilities of CAPT. However, our goal was not comparing autonomous systems with human instruction, but assessing the range of improvement procured by the use of the tools. Finally, although the focus of this thesis has been pronunciation improvement at the segmental level, its combination with the suprasegmental level and other focuses, such as intelligibility or fluency, must be also considered for acquiring a complete pronunciation competence.

4. Scientific Contributions

The scientific impact of this Ph.D. Thesis was supported by sixteen publications at the time of the dissertation presentation. In particular, game-based results of the Alpha experiment have been partially published in [15]. Performance–related results of the TipTopTalk! prototype have been partially published in [16, 20, 21]; whereas results related to the gamification elements included in the CAPT tool have been published in [22, 23]. Regarding the latest game-based prototype, COP, the main results about performance, motivation, and pronunciation improvement have been published in [17] (Journal Citation Reports, JCR Q1).

On the other hand, guided-based experimental results of the English Vowels prototype have been partially published in [24] (JCR Q2) and in [25, 26]. Results of pronunciation improvement and tool description of the Japañol prototype have been partially published in [18, 27, 28]. Other contributions related to this prototype have been partially published in [29, 19]. Also, first results of the latest and ongoing project, Estoñol, have been published in [30] (JCR Q2).

Finally, as a consequence of this work, we have attended nine international conferences and workshops, registered one software program (intellectual property), carried out two predoctoral research stays, obtained two predoctoral fellowships, participated in three research projects, achieved three awards, and developed six CAPT software applications.

5. Acknowledgments

Special thanks to Dr. Valentín Cardeñoso-Payo and Dr. David Escudero-Mancebo, the supervisors of this Ph.D. Thesis; and Dr. César González-Ferreras, Enrique Cámara-Arenas, and Dr. Mario Corrales-Astorgano from the ECA-SIMM research group. We would also like to thank Dr. María Jesús Rodríguez-Triana and the members of the Centre of Excellence in Educational Innovation at the University of Tallinn, Estonia; Mrs. Katrin Leppik and the members of the Institute of Estonian and General Linguistics (University of Tartu, Estonia); Dr. María J. Machuca and the members of Servei de Tractament de la Parla i del So (Autonomous University of Barcelona); and Mr. Takuya Kimura from the Spanish Language and Literature Department (University of Seisen). We truly appreciate the support of the colleagues and researchers of the Department of Computer Science at University of Valladolid, and the financial support of this University for carrying out this work. Also, our sincere thanks and appreciation for the reviewers and the examination board of this Ph.D. Thesis. Finally, we would like to thank the participation of more than five hundred learners in the experiments with the help of companies, language learning centers, and other researchers from all around the world.

6. References

[1] O'Brien *et al.*, "Directions for the future of technology in pronunciation research and teaching," *J. Second Lang. Pronunciation*, vol. 4, no. 2, pp. 182–207, Feb. 2018. [Online]. Available: https://doi.org/10.1075/jslp.17001.obr

[2] R. I. Thomson and T. M. Derwing, "The effectiveness of L2 pronunciation instruction: A narrative review," *Appl. Linguistics*, vol. 36, no. 3, pp. 326–344, Jul. 2015. [Online]. Available: https://doi.org/10.1093/applin/amu076

[3] B. Penning de Vries, C. Cucchiarini, H. Strik, and R. van Hout, "The Role of Corrective Feedback in Second Language Learning: New Research Possibilities by Combining CALL and Speech

Technology," in *Proc. SLATE*, Tokyo, Japan, Sep. 22–24, 2010, pp. 125–130.

[4] G. Seed and J. Xu, "Integrating technology with language assessment: Automated speaking assessment," in *Proc. ALTE*, Bologna, Italy, May 3–5, 2017, pp. 286–291.

[5] C. Tejedor-García, "Design and Evaluation of Mobile Computer-Assisted Pronunciation Training Tools for Second Language Learning," Ph.D. dissertation, Dept. Comput. Sci., Univ. Valladolid, Valladolid, Spain, Sep. 2020. [Online]. Available: https://doi.org/10.35376/10324/43663

[6] Wikipedia contributors, "List of languages by total number of speakers — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=List_of_languages_by_total_number_of_speakers, 2020, [Online; accessed 30-December-2020].

[7] K. Beatty, *Teaching & Researching: Computer-Assisted Language Learning*. Routledge, 2013. [Online]. Available: https://doi.org/10.4324/9781315833774

[8] H. D. Brown and H. Lee, *Teaching by principles: An interactive approach to language pedagogy*. New York, NY, USA: Pearson Education, 2015, vol. 1. [Online]. Available: https://doi.org/10.2307/3587655

[9] Cheon, Jongpil and Lee, Sangno and Crooks, Steven M. and Song, Jaeki, "An investigation of mobile learning readiness in higher education based on the theory of planned behavior," *Comput. Educ.*, vol. 59, no. 3, pp. 1054–1064, Nov. 2012. [Online]. Available: https://doi.org/10.1016/j.compedu.2012.04.015

[10] M. Meeker, "Internet trends 2017," may 2017, Kleiner Perkins, Los Angeles, CA, USA, Rep. [Online]. Available: https://www.bondcap.com/report/it17.

[11] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, Alberta, AB, Canada, Apr. 15–20, 2018, pp. 4779–4783. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461368

[12] A. Neri, C. Cucchiarini, and H. Strik, "The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2–Dutch," *ReCALL*, vol. 20, no. 2, pp. 225–243, May 2008. [Online]. Available: https://doi.org/10.1017/S0958344008000724

[13] T. Greitemeyer and D. O. Mügge, "Video Games Do Affect Social Outcomes: A Meta-Analytic Review of the Effects of Violent and Prosocial Video Game Play," *Personality Social Psychol. Bulletin*, vol. 40, no. 5, pp. 578–589, 2014. [Online]. Available: https://doi.org/10.1177/0146167213520459

[14] D. T. Campbell and J. C. Stanley, *Experimental and quasi-experimental designs for research*. Ravenio Books, 2015.

[15] D. Escudero-Mancebo, E. Cámara-Arenas, C. Tejedor-García, C. González-Ferreras, and V. Cardeñoso-Payo, "Implementation and test of a serious game based on minimal pairs for pronunciation training," in *Proc. SLaTE*, Leipzig, Germany, Sep. 4–5, 2015, pp. 125–130.

[16] C. Tejedor-García, V. Cardeñoso-Payo, E. Cámara-Arenas, C. González-Ferreras, and D. Escudero-Mancebo, "Measuring pronunciation improvement in users of CAPT tool TipTopTalk!" in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 8–12, 2016, pp. 1178–1179.

[17] C. Tejedor-García, D. Escudero-Mancebo, V. Cardeñoso-Payo, and C. González-Ferreras, "Using challenges to enhance a learning game for pronunciation training of English as a second language," *IEEE Access*, vol. 8, no. 1, pp. 74 250–74 266, Apr. 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.2988406

[18] C. Tejedor-García, V. Cardeñoso-Payo, M. J. Machuca, D. Escudero-Mancebo, A. Ríos, and T. Kimura, "Improving Pronunciation of Spanish as a Foreign Language for L1 Japanese Speakers with Japañol CAPT Tool," in *Proc. IberSPEECH*, Barcelona, Spain, Nov. 21–23, 2018, pp. 97–101. [Online]. Available: https://doi.org/10.21437/IberSPEECH.2018-21

[19] C. Tejedor-García and D. Escudero-Mancebo, "Uso de pares mínimos en herramientas para la práctica de la pronunciación del español como lengua extranjera," *Revista de la Asociación Europea de Profesores de Español. El español por el mundo*, vol. 1, no. 1, pp. 355–363, Jan. 2018.

[20] C. Tejedor-García, D. Escudero-Mancebo, E. Cámara-Arenas, C. González-Ferreras, and V. Cardeñoso-Payo, "TipTopTalk! mobile application for speech training using minimal pairs and gamification," in *Proc. IberSPEECH*, Lisbon, Portugal, Nov. 23–25, 2016, pp. 425–432.

[21] C. Tejedor-García, D. Escudero-Mancebo, C. González-Ferreras, E. Cámara-Arenas, and V. Cardeñoso-Payo, "Improving L2 production with a gamified computer-assisted pronunciation training tool, TipTopTalk!" in *Proc. IberSPEECH*, Lisbon, Portugal, Nov. 23–25, 2016, pp. 177–186.

[22] C. Tejedor-García, V. Cardeñoso-Payo, E. Cámara-Arenas, C. González-Ferreras, and D. Escudero-Mancebo, "Playing around minimal pairs to improve pronunciation training," in *Proc. IFCASL*, ser. Feedback in Pronunciation Training Workshop, Saarland, Germany, Nov. 5–6, 2015.

[23] A. Rauber, C. Tejedor-García, V. Cardeñoso-Payo, E. Cámara-Arenas, C. González-Ferreras, D. Escudero-Mancebo, and A. Rato, "TipTopTalk!: A game to improve the perception and production of L2 sounds," in *Abstr. New Sounds 8th Int. Conf. Second Lang. Speech*, Aarhus Univ., Aarhus, Denmark, Jun. 10–12, 2016, p. 160.

[24] C. Tejedor-García, D. Escudero-Mancebo, E. Cámara-Arenas, C. González-Ferreras, and V. Cardeñoso-Payo, "Assessing pronunciation improvement in students of English using a controlled computer-assisted pronunciation tool," *IEEE Trans. Learn. Technol.*, vol. 13, no. 2, pp. 269–282, Mar. 2020. [Online]. Available: https://doi.org/10.1109/TLT.2020.2980261

[25] C. Tejedor-Garcia, V. Cardeñoso-Payo, and D. Escudero-Mancebo, "Design and evaluation of two mobile computer-assisted pronunciation training tools to favor autonomous pronunciation training of english as a foreign language," in *Proc. EDULEARN20*. IATED, 6-7 July, 2020 2020, pp. 7639–7646. [Online]. Available: http://dx.doi.org/10.21125/edulearn.2020.1936

[26] C. Tejedor-García, D. Escudero-Mancebo, C. González-Ferreras, E. Cámara-Arenas, and V. Cardeñoso-Payo, "Evaluating the efficiency of synthetic voice for providing corrective feedback in a pronunciation training tool based on minimal pairs," in *Proc. SLaTE*, Stockholm, Sweden, Aug. 25–26, 2017, pp. 26–30. [Online]. Available: https://doi.org/10.21437/SLaTE.2017-5

[27] C. Tejedor-García, V. Cardeñoso-Payo, and D. Escudero-Mancebo, "Japañol: a mobile application to help improving Spanish pronunciation by Japanese native speakers," in *Proc. IberSPEECH*, Barcelona, Spain, Nov. 2018, pp. 157–158.

[28] T. Kimura, C. Tejedor-García, V. Cardeñoso-Payo, M. J. Machuca, D. Escudero-Mancebo, and A. Ríos, "Japañol, a Computer Assisted Pronunciation Tool for Japanese Students of Spanish Based on Minimal Pairs," in *Abstr. 2nd Int. Symp. Appl. Phonetics*, Aizu, Japan, Sep. 21–23, 2018.

[29] C. Tejedor-García, "Design and Evaluation of a Mobile Application for Second Language Pronunciation Training based on Minimal Pairs," in *Proc. SEPLN 2018*, Seville, Spain, Sep. 21–23, 2018, pp. 7–11. [Online]. Available: http://ceur-ws.org/Vol-2251/paper2.pdf

[30] K. Leppik and C. Tejedor-García, "Estoñol, a computer-assisted pronunciation training tool for Spanish L1 speakers to improve the pronunciation and perception of Estonian vowels," *J. Estonian Finno-Ugric Linguistics (ESUKA – JEFUL)*, vol. 10, no. 1, pp. 89–104, Nov. 2019. [Online]. Available: https://doi.org/10.12697/jeful.2019.10.1.05

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

New tools for the differential evaluation of Parkinson's disease using voice and speech processing

Laureano Moro-Velazquez[1,2], *Jorge A. Gomez-Garcia*[2], *Najim Dehak*[1], *Juan I. Godino-Llorente*[2]

[1]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA.
[2]Bioengineering and Optoelectronics lab (ByO), Universidad Politécnica de Madrid, Madrid, Spain.
laureano@jhu.edu

Abstract

Parkinson's Disease (PD) is a neurodegenerative condition that affects the motor capabilities of individuals. Early detection can potentially contribute to slow its progression in a near future. Therefore, new objective and reliable tools are needed to support its diagnosis. Literature suggests that the patients' speech can provide relevant information about the presence of the disease. In this study, five sets of experiments were carried out, each containing new approaches to detect the presence of the disease in the speech of idiopathic PD patients and control speakers from three parkinsonian corpora, two of them in the Spanish language. Different speech frame selection techniques are proposed, such as phonemic and acoustic landmark distillation, providing certain specific speech segments of interest to this work's purposes. Multiple cepstral and spectral features were employed, along with several classification techniques based on Gaussian models and speaker embeddings. The best accuracy results in detecting PD with the proposed methodologies reached values ranging from 85% to 94% with Area Under the Curve between 0.91 and 0.99, depending on the corpus. Results suggest that PD affects the movements related to all of the studied articulatory segmental groups but has a more evident influence in the consonants with a greater narrowing of the vocal tract, mainly plosives and fricatives. The new proposed methodologies demonstrate their ability to support PD's diagnosis during a patient's clinical assessment and are a step forward in PD's speech-based diagnosis systems.

Index Terms: Parkinson's Disease, speaker recognition, GMM-UBM, phonemic distillation, acoustic landmarks

1. Introduction

In this document we present an overview of the thesis [1] of the first author titled "Towards the differential evaluation of Parkinson's Disease by means of voice and speech processing"[1] to the Iberspeech 2020 *Ph. D. Thesis Special Session*.

Parkinson's Disease (PD) affects to 1% of the population over the age of 60 in industrialized countries and, with increasing life expectancy, it is expected to affect to more than 9 million people in industrialized nations by 2030 [2]. A slower advance of the disease in patients will diminish its impact on their daily activities and increase their quality of life. It will also greatly reduce the economic burden of PD of health-care systems. Unfortunately, the diagnosis is based on the assessment of a patient's signs and symptoms over an observation period which can last from months to years. Since PD is a motor system disorder, the analysis of a patient's movements while performing a complex motor task can lead to the identification of potential biomarkers. But, which type of tasks can be used within the analysis? Multiple studies have found enough evidences to propose biomarkers or automatic diagnosis schemes based on a patient's eye movements [3], or handwriting [4], among others. Speech can also be used to evaluate PD because it involves coordination and precision of movements in mainly the laryngeal and articulatory muscles [5].

Multiple studies have analyzed the phonatory, articulatory, prosodic and linguistic aspects of parkinsonian speech, and of them have proposed the use of artificial intelligence to provide diagnostic and assessment tools. Typical approaches include voice quality measurements (noise, frequency and amplitude perturbations, etc.) and classifiers such as Support Vector Machine (SVM) or Random Forest to identify patients with PD employing sustained phonations [6, 7]. Others propose the use of speech characterizations such as filterbank features or Mel-Frequency Cepstral Coefficients (MFCC) as input to SVM, Gaussian Mixture Model (GMM) and, in some cases, neural networks [8, 9]. An extensive review of these approaches and a list of common features and corpora can be found in [10, 11].

The purpose of this study is to propose new approaches to support the clinical diagnosis of PD by employing speech as the object of observation and to obtain new information about the influence of PD in different articulatory movements.

2. Hypothesis and goals

The phonatory and articulatory aspects of the patients with idiopathic PD are affected by the motor dysfunctions associated with this neurodegenerative disease [5]. Our hypothesis is that the resulting speech impairments, inaccuracy of articulatory movements and modified patterns of acceleration and velocity of the articulators can be characterized using signal processing techniques. This characterization, combined with machine learning classification schemes employed in speaker and speech recognition, can yield new tools for the diagnosis of PD.

The main goal of this study is to analyze distinct advanced speech processing technologies and machine learning techniques for the detection and assessment of PD from the speech's articulatory and phonatory aspects. Some specific objectives derived from the main goal are: 1) the analysis of supervised schemes to detect PD from phonatory aspects of speech; 2) the analysis of several supervised and unsupervised new schemes to detect PD from articulatory aspects of speech; 3) the analysis of the role of the distinct phones and articulatory manners (plosives, fricatives, nasals and liquids) in the automatic detection of PD; 4) the analysis of the distinct transitions between phones and their influence in the detection of PD; 5) the identification of the advantages and disadvantages of the different speech tasks in the proposed schemes to detect PD; 6) the analysis of the combination of a phonatory and articulatory subsystems to au-

[1]Available at:
http://oa.upm.es/51278/1/LAUREANO_MORO_VELAZQUEZ.pdf

10.21437/IberSPEECH.2021-36

tomatically detect PD; 7) the study of articulators' kinematics and its relevance in the automatic detection of PD; 8) The study of the generalization properties of the proposed systems.

3. Methodology

In this study we propose new approaches for the automatic detection of PD employing speech technologies which had not been exploited before for this task. Additionally, a combination of the articulatory and phonatory aspects is explored with the aim of using the potential complementary capabilities of the correspondent used methodologies for the automatic detection of the disease. Within each experiment, different approaches or variations of the same basis scheme were proposed in order to find the optimal solution. Six speech corpora, described in Section 4, were employed to carry out the experiments.

3.1. Experimental set 1: Speaker recognition technologies

In this experiment [12], several speaker recognition techniques were applied and adapted for the automatic detection of PD using the patient's speech. Three families of features were considered, MFCC, Rasta-Perceptual Linear Prediction (PLP) and Linear Predictive Coding (LPC), along with their respective derivatives, utilizing multiple configurations. Equally, two classification techniques, namely Gaussian Mixture Model-Universal Background Model (GMM-UBM) and i-Vectors with Gaussian Probability Linear Discriminant Analysis (GPLDA), were used to train and test the automatic detectors. The objective of this study was twofold: firstly to evaluate the application of these techniques to a new scenario, analyzing their different degrees of freedom to establish a baseline to compare results with further studies; and secondly, to evaluate the influence of kinetic changes of instantaneous coefficients and the importance of the time window used to estimate the derivatives for the detection of PD. A large amount of trials were performed using a single parkinsonian and a single auxiliary corpus. The configuration leading to the best results was tested again with the other two remaining parkinsonian corpora. Finally, cross-corpora trials were performed to validate the methodology at the optimum configuration.

3.2. Experimental set 2: Forced Gaussians

In this second experiment [13], several approaches using GMM are proposed. In this experiment, a Forced Alignment Model (FAM) was built to automatically segment and align all the frames on the speech in the parkinsonian and Universal Background Model (UBM) corpora. The resulting phonetic labels were employed to build GMM-UBM models containing Gaussians which were specific for each phonetic label (forced-GMM), yielding models able to compare all the phonetic units of the parkinsonian and control speakers more precisely. The different trials were performed in the three parkinsonian corpora. Finally, a group of cross-corpora trials was performed to validate the optimal configurations. These experiments allowed us to obtain more precise PD detectors and to evaluate which phones were more relevant in the automatic detection of PD.

3.3. Experimental set 3: Phonemic distillation

As several works in the literature point out, the consonants produced after a strong constriction of the articulators are usually more affected by PD [14, 15, 16]. The idea behind the third experiment [17, 18] was to use only specific segments of the speech, depending on the manner of articulation and the narrowing of the vocal tract to analyze their influence in the detection of PD. The speech signals from the different corpora were segmented and labeled (when possible) by using forced alignment techniques. The segments containing a single phone category depending on the manner of articulation (fricative, liquid, nasal, plosive and vowels) were used to train and test separately GMM-UBM classifiers (one classifier per manner) employing PLP as features. Thus, the obtained models to detect PD were focused on these different phonetic groups. The different trials were performed in three parkinsonian corpora. In this case, a fusion of scores coming from the models trained using different specific segments was also carried out. Finally, a group of cross-corpora trials was performed to validate the optimal configurations.

3.4. Experimental set 4: Acoustic landmark distillation

In this fourth articulatory experiment [19], different speech segments related to relevant articulatory moments, such as bursts, transitions between vowels and consonants or the beginning and end of glottal activity, were used to identify relevant transition segments which were employed to detect PD, using an acoustic landmark distillation. This experiment is similar to Experiment 3 but in this case, instead of using whole phonetic segments, only the transitions are used. To identify the acoustic landmarks and, thus, the transition segments, the methodology proposed by [20] was followed. In this case, new GMM-UBM classifiers were trained and tested using *burst*, *sonorant*, and *glottal* transition segments separately. Then, a fusion of scores from

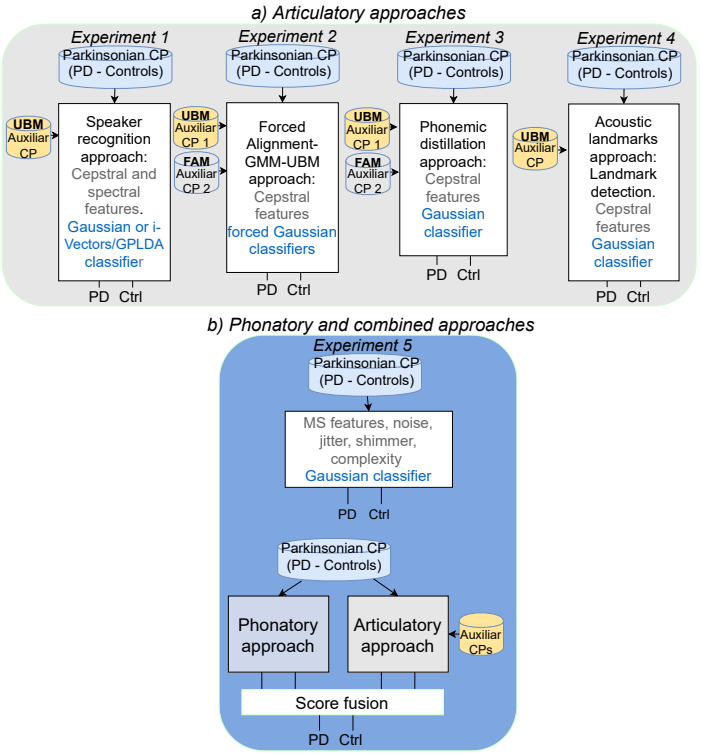


Figure 1: *Overview of the methodology employed in the study, including the four experiments on the articulatory approaches (a), the phonatory approaches, and the combination of the phonatory approach with one articulatory approach (b), where the articulatory part corresponds to the methodology providing the best results among the four articulatory experiments. CP stands for 'corpus'.*

the different models (one per segment type) was performed to combine systems.

3.5. Experimental set 5: Phonatory analysis and score fusion

A large portion of the approaches found in literature use jitter, shimmer and noise to detect PD [11] but not many of them propose new features for this task. In the experimental set 5 [21], a new group of features based on the use of Modulation Spectrum (MS) [22, 23][2] were used in a GMM classification scheme to automatically discriminate patients from controls in three different corpora. These features, containing information about noise, voice frequency and amplitude perturbations, and tremor, were considered as appropriate for the proposed task. To finish the experiments, a fusion of scores between the phonatory approach and the articulatory approach yielding the best results in terms of accuracy and Area Under the Curve (AUC) was performed in order to combine the two aspects into a single approach.

A scheme of all of the experiments studied in this study and the final combination of aspects can be found in Figure 1.

4. Corpora

In this study, six corpora were employed: Neurovoz [13], GITA [24], CzechPD [25], Albayzin [26], Vystadial-Czech [27] and FisherSP [28]. The first three corpora contain different speech tasks from parkinsonian and control speakers and were used to train and test the models proposed in the methodology. The three latter corpora are considered auxiliary and were employed to train UBMs and FAMs. The mother tongue of the subjects in GITA, Neurovoz, Albayzin and FisherSp is the Spanish language, whereas CzechPD and Vystadial-Czech contain the speech from Czech speakers.

The Neurovoz corpus[3] is a new speech dataset recorded in this study. It contains 47 parkinsonian and 32 control speakers whose mother tongue is Spanish Castillian. This corpus was recorded in collaboration with the otorhinolaryngology and neurology departments of the Gregorio Marañón hospital in Madrid, Spain.

5. Results and discussion

Table 1 contains the best results of the five experimental sets for the parkinsonian corpora employing different speech tasks: Text-Dependent Utterance (TDU), Diadochokinetic (DDK), monologues and sustained vowel /a:/. Table 2 includes the cross-corpora results for experimental sets 1, 2, 3, and 4. In this last case, each parkinsonian corpus was used to test the models trained with the other two parkinsonian datasets. Best results per corpus are shaded and best overall results are in bold. The first experimental set is considered the baseline of this study and provides maximum accuracy results between 81% and 91%, with AUC between 0.88 and 0.94 depending on the employed corpus and speech task.

The forced-GMM techniques proposed in the experimental set 2 can be considered novel on PD detection, and these allow us to observe the influence of PD in each of the individual phonetic units. The results suggest that phonetic units requiring a higher narrowing of the vocal tract but without a burst tend to be more influential in the detection of PD, while nasal consonants are the less influential. The results suggest that the phones /C/, /D/, /g/ and /R/ tend to produce better accuracy, and /m/ and /B/ are less relevant in the detection of PD.

Additionally, other two new techniques were proposed in this study: phonemic distillation (experimental set 3) and acoustic landmark distillation (experimental set 4). While both techniques are based on frame selection, the first one uses specific speech segments related to the manner of articulation, such as plosives or fricatives, while the second one is focused in the selection of certain types of transitions (from silence to plosive phone or from vowel to consonants). The use of phonemic or landmark distillation in the UBM produces improvements respect to the baseline (Experimental set 1) in the three corpora and with almost any speech task. Attending to Tables 1 and 2, the results of Experimental set 3 (phonemic distillation) tend to outperform the rest of the approaches in the k-folds cross-validations and in the cross-corpora validations. Cross-validation results are reported at speaker level. In the cross-validation process, the speakers considered in the training folds were not included in the testing folds.

Concerning the cross-corpora tests (carried out using only

Table 1: *Best cross-validation results for each experiment as a function of the employed corpus and speech task. Sust. v. stands for sustained vowel, and Exp. for Experimental set.*

Corpus	Speech task	Exp.	Accuracy ± CI	AUC	Sens.	Spec.
GITA	TDU	1	80 ± 8	0.85	0.82	0.78
		2	81 ± 8	0.88	0.84	0.78
		3	$\mathbf{85 \pm 7}$	**0.91**	0.82	0.88
		4	85 ± 7	0.89	0.84	0.86
	DDK	1	81 ± 8	0.88	0.82	0.8
		2	79 ± 8	0.86	0.86	0.72
		3	83 ± 7	0.89	0.86	0.8
		4	83 ± 7	0.88	0.86	0.8
	Monol.	1	80 ± 8	0.88	0.76	0.84
		2	78 ± 8	0.84	0.73	0.82
		3	82 ± 8	0.89	0.8	0.84
		4	80 ± 8	0.86	0.76	0.84
	Sust. v.	5	71 ± 9	0.8	0.72	0.7
	TDU + Sust. v.	5	85 ± 7	0.91	0.82	0.88
Neurovoz	TDU	1	86 ± 8	0.93	0.87	0.84
		2	81 ± 9	0.87	0.83	0.78
		3	$\mathbf{89 \pm 7}$	**0.93**	0.87	0.91
		4	$\mathbf{89 \pm 7}$	**0.93**	0.91	0.84
	DDK	1	79 ± 9	0.85	0.87	0.65
		2	79 ± 8	0.86	0.86	0.72
		3	86 ± 8	0.88	0.89	0.81
		4	83 ± 7	0.88	0.86	0.8
	Monol.	1	79 ± 12	0.81	0.59	0.9
		2	66 ± 14	0.67	0.35	0.83
		3	77 ± 12	0.79	0.53	0.9
		4	79 ± 12	0.9	0.47	0.97
	Sust. v.	5	64 ± 10	0.68	0.75	0.48
	TDU + Sust. v.	5	87 ± 7	0.94	0.85	0.91
CzechPD	DDK	1	88 ± 1	0.94	0.85	0.93
		2	94 ± 1	0.97	0.9	1
		3	94 ± 1	0.98	0.9	1
		4	$\mathbf{94 \pm 1}$	**0.99**	0.9	1

[2]MS features code available at: https://github.com/jorgomezga/AVCA-ByO

[3]Data and more information about this corpus are available at https://zenodo.org/record/3557758

Table 2: *Best cross-corpora results for all the experiments. In all cases, speech task is DDK.*

Test corpus	Exp.	Accuracy ± CI	AUC	Sens.	Spec.
GITA	1	73 ± 9	0.82	0.84	0.62
	2	66 ± 9	0.76	0.9	0.42
	3	75 ± 8	0.84	0.86	0.64
	4	73 ± 9	0.81	0.8	0.66
Neurovoz	1	75 ± 10	0.82	0.8	0.65
	2	74 ± 10	0.78	0.87	0.5
	3	81 ± 9	0.83	0.91	0.62
	4	76 ± 10	0.83	0.78	0.73
CzechPD	1	79 ± 14	0.91	1	0.5
	2	76 ± 14	0.87	0.95	0.5
	3	82 ± 13	0.95	0.85	0.79
	4	82 ± 13	0.95	1	0.57

DDK task as it is the only common task in the three corpora) the best AUC is never reduced more than 0.05 absolute points respect to the k-folds trials, as it is observed in Table 2. The use of cross-corpora trials –i. e. classifiers trained with a corpus and evaluated with a different one– for the automatic detection of PD is a novelty by itself since, to the authors of this study's knowledge, no published work had succesfully performed this type of trials before.

Globally, the highest accuracy obtained in the trials is 94% with AUC of 0.99, sensitivity of 0.90 and specificity of 1.00, in the CzechPD corpus using acoustic landmark distillation on the UBM, Experimental set 4. It is important to consider that this corpus mainly contains newly diagnosed patients. Therefore, results suggest that the new proposed approaches can be valid detecting PD in early stages. The best results obtained in the k-folds cross-validation in the three corpora are always over 85%, as shown in Table 1.

In this study four types of speech tasks were considered: TDU, DDK tasks, monologues and sustained vowel /a/. TDU provided the best results in almost every trial, as it can be inferred from Table 1. In this respect, a detailed analysis of the results reveals that the impact of PD on speech is not limited to a few articulatory movements, phones or transitions, but influences them all in a higher or lower degree. TDU contain more variety of phones and articulatory movements than DDK tasks, justifying the differences between the results provided by both tasks. On the other hand, although monologues include a larger variety of articulatory movements than DDK, these are different for each utterance in the training and testing sets and the resulting models are text-independent. Text-dependent models demonstrate to outperform text-independent approaches as the first ones allow to compare more precisely the same specific segments (transitions, phones, acceleration or velocity of articulatory movements, etc).

These experiments were carried out employing only idiopathic PD patients and controls, and no other parkinsonism or neurodegenerative disease such as Friedrich's ataxia or Huntington's Disease (HD) has been considered. Consequently, it is not possible to evaluate if the proposed approaches provide discrimination between idiopathic PD and other neurodegenerative diseases, important for early diagnosis, which will be carried out in future studies. The analysis of results from male and female subjects separately, creating their respective specific models would have been desirable. However, larger corpora are advisable for these purposes.

Finally, the analysis performed in this work has been complemented in further studies employing other state-of-the-art speech and speaker recognition methodologies such as end-to-end Automatic Speech Recognition (ASR) models [29], x-vectors [30] or Long Short-Term Memory (LSTM) networks [31], providing new approaches and insights about how PD affects the speech of patients.

6. Conclusions

In this study, new approaches to support the differential evaluation of PD by means of voice and speech processing have been proposed and analyzed.

One of the first conclusions obtained from the analysis of results is that kinematic changes, characterized by means of the derivatives of PLP coefficients of speech are crucial in the detection of PD using speech. These provide information about acceleration and velocity of the articulators during speech, two characteristics that are highly influenced by the disease in the forms of hypokinesia or other types of kinetic impairments such akinesia or bradykinesia.

Also, since the motor dysfunctions caused by PD produce misarticulation, results suggest that these deficits can be found in different types of segments of speech (i.e, distinct groups of phones and transitions). The articulatory movements requiring a higher narrowing of the vocal tract, that normally occur during the pronunciation of fricatives and plosives, are the most influenced by the disease. However, these are not the only segments of speech containing information about PD since schemes based only on vowels provide significant results too. Results suggest that PD influences all the studied phonetic groups, affecting to the articulatory sequence as a whole. For this reason, the analysis of phonetically balanced speech tasks allows to evaluate the presence of PD from speech by using automatic detectors. Additionally, when employing TDU as speech tasks, the obtained classification models are text-dependent and allow to compare more precisely the articulation of patients and controls since all the speakers repeat the same sequence of phonemes.

Regarding the two studied speech aspects, the discriminatory properties of the proposed phonatory approaches are quite limited in comparison with the articulatory approaches. The combination of these approaches yields better results in some cases but the advantages of this combination are unclear.

Finally, the best results obtained with the proposed methodologies reach accuracy values ranging from 85% to 94% with AUC between 0.91 and 0.99 and sensitivity between 0.82 and 0.91 depending on the parkinsonian corpus considered. These results are obtained employing the proposed phonemic distillation technique in the UBM corpus in a GMM-UBM classification scheme. These values can be considered close to the maximum feasible accuracy which is estimated to be around 95% due to possible misdiagnosis in the corpora [13]. These true limits have not been considered before in other works performing automatic detection of PD. The proposed approaches exhibit a lower accuracy in the cross-corpora trials with respect to the cross-validation (k-folds) trials but still between 75% and 82% with AUC between 0.84 and 0.95 and sensitivity between 0.85 and 1, depending on the testing corpus, indicating that these approaches generalize in scenarios with different recording conditions.

7. Acknowledgements

This work was supported by the Government of Spain under grants TEC2012-38630-C04-01 and DPI2017-83405-R1 and by the MISTI Global Seed Funds Award ("Objective Evaluation of PD and other parkinsonisms from the speech" project).

8. References

[1] L. Moro-Velázquez, "Towards the differential evaluation of parkinson's disease by means of voice and speech processing," Ph.D. dissertation, 2018. [Online]. Available: http://oa.upm.es/51278/1/LAUREANO_MORO_VELAZQUEZ.pdf

[2] E. R. Dorsey *et al.*, "Projected number of people with parkinson disease in the most populous nations, 2005 through 2030," *Neurology*, vol. 68, no. 5, pp. 384–386, 2007.

[3] F. Chan, I. T. Armstrong, G. Pari, R. J. Riopelle, and D. P. Munoz, "Deficits in saccadic eye-movement control in parkinson's disease," *Neuropsychologia*, vol. 43, no. 5, pp. 784–796, 2005.

[4] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zanuy, "Decision support framework for parkinson's disease based on novel handwriting markers," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 23, no. 3, pp. 508–516, 2015.

[5] J. R. Duffy, *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2013.

[6] J. Mekyska, Z. Galaz, Z. Mzourek, Z. Smekal, I. Rektorova, I. Eliasova, M. Kostalova, M. Mrackova, D. Berankova, M. Faundez-Zanuy *et al.*, "Assessing progress of parkinson's disease using acoustic analysis of phonation," in *Bioinspired Intelligence (IWOBI), 2015 4th International Work Conference on*. IEEE, 2015, pp. 111–118.

[7] J. Rusz, J. Hlavnička, T. Tykalová, M. Novotný, P. Dušek, K. Šonka, and E. Růžička, "Smartphone allows capture of speech abnormalities associated with high risk of developing parkinson's disease," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 8, pp. 1495–1507, 2018.

[8] M. Novotny, J. Rusz, R. mejla, and E. Rka, "Automatic evaluation of articulatory disorders in parkinson s disease," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1366–1378, 2014.

[9] J. Vásquez-Correa, T. Arias-Vergara, J. Orozco-Arroyave, and E. Nöth, "A multitask learning approach to assess the dysarthria severity in patients with parkinson's disease," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 456–460, 2018.

[10] L. Moro-Velazquez and N. Dehak, "A review of the use of prosodic aspects of speech for the automatic detection and assessment of parkinson's disease," *Automatic Assessment of Parkinosnian Speech. Communications in Computer and Information Science (CCIS) Series*, vol. 1295, pp. 53–73, 2020.

[11] L. Moro-Velazquez, J. A. Gomez-Garcia, J. D. Arias-Londoño, N. Dehak, and J. I. Godino-Llorente, "Advances in parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects," *Biomedical Signal Processing and Control*, vol. 66, p. 102418, 2021.

[12] L. Moro-Velazquez, J. A. Gomez-Garcia, J. I. Godino-Llorente, J. Villalba, J. R. Orozco-Arroyave, and N. Dehak, "Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect parkinsonś disease," *Applied Soft Computing*, vol. 62, pp. 649–666, 2018.

[13] L. Moro-Velazquez, J. A. Gomez-Garcia, J. I. Godino-Llorente, J. Villalba, J. Rusz, S. Shattuck-Hufnagel, and N. Dehak, "A forced gaussians based methodology for the differential evaluation of parkinson's disease by means of speech processing," *Biomedical Signal Processing and Control*, vol. 48, pp. 205–220, 2019.

[14] J. A. Logemann and H. B. Fisher, "Vocal Tract Control in Parkinson's Disease," *Journal of Speech and Hearing Disorders*, vol. 46, no. 4, p. 348, 1981.

[15] J. Kegl, H. Cohen, and H. Poizner, "Articulatory consequences of Parkinson's disease: perspectives from two modalities." *Brain and Cognition*, vol. 40, no. 2, pp. 355–86, 1999.

[16] J. Godino-Llorente, S. Shattuck-Hufnagel, J. Choi, L. Moro-Velázquez, and J. Gómez-García, "Towards the identification of idiopathic parkinson's disease from the speech. new articulatory kinetic biomarkers," *PloS one*, vol. 12, no. 12, p. e0189583, 2017.

[17] L. Moro-Velazquez, J. Gomez-Garcia, J. I. Godino-Llorente, J. Rusz, S. Skodda, F. Grandas, J.-M. Velazquez, J. R. Orozco-Arroyave, E. Noth, and N. Dehak, "Study of the automatic detection of parkison's disease based on speaker recognition technologies and allophonic distillation," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 1404–1407.

[18] L. Moro-Velazquez, J. A. Gomez-Garcia, J. I. Godino-Llorente, F. Grandas-Perez, S. Shattuck-Hufnagel, V. Yagüe-Jimenez, and N. Dehak, "Phonetic relevance and phonemic grouping of speech in the automatic detection of parkinson's disease," *Scientific Reports*, vol. 9, no. 1, pp. 1–16, 2019.

[19] L. Moro-Velazquez, J. Godino-Llorente, J. Gómez-García, J. Villalba, S. Shattuck-Hufnagel, and N. Dehak, "Use of acoustic landmarks and gmm-ubm blend in the automatic detection of parkinson's disease," in *Models and Analysis of Vocal Emissions for Biomedical Applications: 10th International Workshop, december, 13-15, 2017*, vol. 117. Firenze University Press, 2017, p. 73.

[20] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1872–1891, 2002.

[21] L. Moro-Velázquez, J. A. Gómez-García, N. Dehak, and J. I. Godino-Llorente, "Analysis of phonatory features for the automatic detection of parkinson's disease in two different corpora," *Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA) 2019*, p. 33, 2019.

[22] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, and G. Andrade-Miranda, "Modulation spectra morphological parameters: A new method to assess voice pathologies according to the grbas scale," *BioMed research international*, vol. 2015, p. 259239, 2015.

[23] L. Moro-Velazquez, J. A. Gomez-Garcia, and J. I. Godino-Llorente, "Voice pathology detection using modulation spectrum-optimized metrics," *Frontiers in bioengineering and biotechnology*, vol. 4, p. 1, 2016.

[24] J. Orozco-Arroyave and J. Arias-Londoño, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease." *Proceedings on the International Conference on Language Resources and Evaluation (LREC)*, pp. 342–347, 2014.

[25] J. Rusz *et al.*, "Imprecise vowel articulation as a potential early marker of parkinson's disease: Effect of speaking task," *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2171–2181, 2013.

[26] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Mariño, and C. Nadeu, "Albayzín speech database: Design of the phonetic corpus," in *Eurospeech 1993. Proceedings of the 3rd European Conference on Speech Communication and Technology*, vol. 1. ISCA, 1993, pp. 175–178.

[27] M. Korvas, O. Platek, O. Dusek, L. Zilka, and F. Jurcicek, "Free english and czech telephone speech corpus shared under the cc-by-sa 3.0 license," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014, pp. 4423—-4428.

[28] "Fisher spanish speech." [Online]. Available: https://catalog.ldc.upenn.edu/LDC2010S01

[29] L. Moro-Velazquez, J. J. Cho, S. Watanabe, M. A. Hasegawa-Johnson, O. Scharengorb, H. Kim, and N. Dehak, "Study of the performance of automatic speech recognition systems in speakers with parkinson's disease," *Interspeech 2019*, pp. 3875–3879, 2019.

[30] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect parkinson's disease from speech," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1155–1159.

[31] S. Bhati, L. Moro-Velazquez, J. Villalba, and N. Dehak, "Lstm siamese network for parkinson's disease detection from speech," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2019, pp. 1–5.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Prosody training of people with Down syndrome using an educational video game

Mario Corrales-Astorgano[1]

[1]Department of Computer Science, University of Valladolid, Spain
mcorrales@infor.uva.es

Abstract

The speech of people with Down syndrome presents multiple disorders affecting the different components of language (syntax, semantics, phonology and pragmatics). In particular, prosody is also affected, conditioning their personal development and their social integration. Due to these difficulties, prosody training is fundamental in their speech therapy. One aim of this work is the definition of a video game focused on training some language skills, specifically those ones related with prosody. Other aim is the creation of an analysis system to evaluate the prosodic quality of users' utterances. The results show that the video game is useful to keep the attention of the players with Down syndrome during all game session. In addition, some statistically significant differences are found between people with Down syndrome and people without intellectual disabilities in the frequency, energy, temporal and spectral domain. The accuracy of identifying a recording as produced by a person with Down syndrome or by a person without intellectual disabilities is up to 95%. Finally, an accuracy of 79.3% is achieved in the task of predicting the prosodic expert evaluation using an automatic classifier trained with the acoustic features extracted from the recordings of people with Down syndrome.

1. Introduction

Some people with Down syndrome (DS) have problems in their social relationships due to their communicative problems [1, 2, 3]. Speech in general [4] and prosody in particular [5] are affected, producing problems in their communicative skills. There are few works in the literature that have analyzed the speech of people with Down syndrome using a corpus comparative approach [4]. The majority of the studies presented in the literature have used a perceptual test approach, while there are few studies based on comparing or analyzing acoustic features extracted from the recordings of the corpus. The recording of a speech corpus of people with Down syndrome is a hard task, because these people present cognitive problems like short time memory problems or attention deficits, among others [1]. In addition, characterizing prosodic impairments in populations with developmental disorders is a hard task [6]. However, prosody assessment procedures appropriate for using with individuals with intellectual and/or developmental disabilities need to be employed, with the aim of being useful for speech therapists in their therapy with people with Down syndrome.

The potential of games to improve motivation and engagement in education has been examined [7]. However, there are few of them focused on speech training of people with Down syndrome [4]. Although there are some tools described in the literature [8], the cognitive problems of people with Down syndrome difficult them to use these tools.

In this work, the main objective was the development of an educational video game focused on improving the communication skills of people with Down syndrome, specifically prosodic skills. To reach this objective, the main communication problems of this population had been analyzed with the aim of developing training activities focused on improving these problems. To improve the motivation of the players, it was important to define the scenarios and the narrative where the training activities were included. The analysis of how the cognitive problems of people with Down syndrome could affect the interaction with the video game had been studied to design an effective interface. In addition, the automatic evaluation of the audios recorded in the training activities could be useful to enhance the autonomous use of the video game by the players with Down syndrome. This was a thesis by compendium of publications, so each of the 3 papers was focused on different objectives of the thesis. The video game design and evaluation was analyzed in [9], the study of differences between acoustic features of the speech of people with Down syndrome and the speech of people without intellectual disabilities was developed in [10] and the automatic evaluation of the recordings of the video game was presented in [11].

The structure of the article is as follows. Section 2 reviews related works from the state of the art. Section 3 describes the characteristics of the video game. Section 4 describes a summary of the different methodologies followed to reach the thesis objectives. Section 5 shows a summary of the results obtained in this thesis related with the video game evaluation, the analysis of the differences between acoustic features of the speech of people with Down syndrome and the speech of people without intellectual disabilities and a first approach to the automatic evaluation of the video game recordings. Finally, section 6 describes the conclusions of this work.

2. Previous work

The are mainly two approaches that have been followed to analyze the prosody problems of people with Down syndrome: acoustic analysis and perceptual analysis. In both cases, the age of the population selected for the study seems to be important for the results obtained, due to the physiological differences between children and adults.

Concerning the acoustic approach, [12, 13, 14] found significantly higher F0 values in adults with Down syndrome as compared to adults with typical development (TD). In addition, [12] found lower jitter (frequency perturbations) in adult speakers with Down syndrome than in TD adults. As for energy, [14] found significantly lower energy values in adults with Down syndrome than in TD people. Moreover, [15] concluded that adults with Down syndrome had poor control over energy in stressed versus unstressed vowels. Finally, temporal domain results depend on the unit of analysis employed. [15] found that people with cognitive disorders presented an excessive variability in vowel duration, while [13] and [16] reported longer du-

10.21437/IberSPEECH.2021-37

rations of vowels in adults with Down syndrome than in TD adults. [14] discovered a lower duration of words in male adults with Down syndrome than male TD adults. Moreover, people with Down syndrome present some disfluency problems. Although disfluency (stuttering or cluttering) has not been demonstrated as a universal characteristic of Down syndrome, it is a common problem of this population [17, 18, 19]. These disfluencies can affect the speech rhythm of people with Down syndrome. On the other hand, [20] indicated that children with Down syndrome had lower F0 than children without intellectual disabilities. [21] found higher jitter in children with Down syndrome than children without intellectual disabilities. In terms of energy, [21] indicated higher shimmer in children with Down syndrome than in children without intellectual disabilities. Perceptual studies show mixed results. [21] described the voice of children with Down syndrome as being statistically different from the voice of children without intellectual disabilities in five speech problems: grade, roughness, breathiness, asthenic speech and strained speech. [22] judged the voice quality of adults with Down syndrome as hoarse. In addition, [23] noted discrepancies between perceptual judgments of pitch level and acoustic measures of F0.

Concerning educational video games, there are some studies that show the efficiency of ICT and video games in the cognitive rehabilitation and teaching of people with intellectual disability: improvement in choice reaction time [24], stimulating cognitive abilities of children [25], and independent decision making [26]. Furthermore, other games are focused on therapy for children with speech disorders [27]. Reviews of educational software for people with DS can be found in [28].

Automatic assessment of pathological speech has also been researched, but, in general, the studies on the topic are related to specific aspects and populations. Some works focus on the speech intelligibility of people with aphasia [29] or speech intelligibility in pathological voices [30]. Others try to identify speech disorders in children with cleft lip and palate [31] or to predict automatically some dysarthric speech evaluation metrics, such as intelligibility, severity and articulation impairment [32]. All these works include a subjective evaluation carried out by experts as a reference to train the classification systems.

3. Game description

The video game, named as PRADIA, has the structure of a graphic adventure game, including conversations with characters, getting and using items and navigating through scenarios (Figure 1). Players have to use the mouse to interact with the elements of the game. Players go through different scenarios where they have to do some actions, like solving an activity or using an item. Some activities focus on lexical-semantic comprehension and on improving of prosodic perception in specific contexts. Other activities focus on oral production, so the player is encouraged by the game to train his speech, keeping in mind prosodic aspects like intonation, expression of emotions or syllabic emphasis. In the activities, the player is introduced by the game into different conversations with game characters, where the player has to choose between different options to continue the dialogue (Comprehension activities) or to record some sentences related with the dialogue context (Production activities), depending on the activity. Finally, there are other activities that were included to add variety to the game and to train other skills not directly related to speech training (Visual activities).

During the game session, information about user interaction is stored, as well as the audio recordings of the production

Figure 1: *Production activity inside a game scenario*

activities. This information was used to compare the interaction between people with Down syndrome and people without intellectual disabilities. Additionally, the audio recordings increase the speech corpus. This user interaction log has information about game and activities duration, the attempts to complete a task, number of mouse clicks or the helps showed to the user.

4. Methodology

Figure 2 shows the methodology followed to design, develop and evaluate the video game. This is a user-centered design methodology, where Down syndrome experts, therapist and final users participated in the game development. After the development, the game was evaluated by people with Down syndrome taken into account the five usability aspects defined in [33] (easy to learn, effective, efficient, engaging and error tolerant). These evaluations were carried out with 14 users with Down syndrome (10 boys and 4 girls, chronological age between 13 and 39 years), 10 children without intellectual disabilities and 10 adults without intellectual disabilities. To analyze the differences in the game interaction among groups, objective and subjective evaluation methods were combined. On the one hand, the game itself recorded data about the interaction between the player and the game. On the other hand, at the end of the game session, evaluators gave a questionnaire on general aspects of usability to the player; this was complemented by observations that evaluators collected during the test. In addition, at the end of the questionnaire, the opinions of speech therapists or teachers who are dedicated to special education were collected. As the video game records the voice of the players, we used these recordings in a perception test, to check whether the players' pronunciation improves as they use the video game.

Figure 3 shows the experimental methodology followed to identify the differences of the acoustic and prosodic features of people with Down syndrome in comparison with people without intellectual disabilities. Firstly, the speech corpus recorded by people with Down syndrome and by typically developing people was gathered (349 recordings of 18 speakers with Down syndrome and 250 recordings of 22 speakers without intellectual disabilities). Secondly, acoustic features were extracted from all the recordings of each corpus using the openSmile software [34] and the feature set GeMAPS [35]. Thirdly, a non-parametric statistical test to analyze the differences between groups was carried out. Fourthly, the automatic classification experiment was carried out, in which the features with signifi-

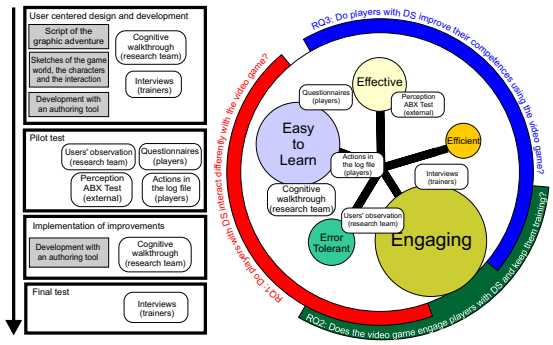


Figure 2: *Evaluation strategy. The circles represent the usability aspects (the size reflects the importance of the aspect), the rectangles represent the development and evaluation phases, and the rounded rectangles are the evaluation activities with the participants in brackets.*

cant differences were used to train three binary classifiers. The performance of these classifiers was obtained using 10-folds cross validation, where 90% of the data was used to train the classifier and 10% was used to test the classifier, repeating this process 10 times. To analyze the performance of the classification, we used the classification rate. The unweighted average recall (UAR) was also used. This metric is the mean of sensitivity (recall of positive instances) and specificity (recall of negative instances). UAR was chosen as the classification metric because it equally weights each class regardless of its number of samples, so it represents more precisely the accuracy of a classification test using unbalanced data. Finally, in order to evaluate the impact of prosody in the perception of the listeners, prosody transfer techniques were used. These techniques consist of transfers, phoneme by phoneme, of the pitch, energy and duration from one audio to another. Therefore, the new audio file contains the original utterance but with the prosody transferred from another utterance. This process was applied to some audios of people with Down syndrome and other audios from people without intellectual disabilities. Afterwards, some people without specific knowledge in prosody evaluated each modified recording using a 5-point Likert scale, indicating the security degree in which they identify each recording as belonging to a person with an intellectual disability.

Figure 4 describes the experimental procedure followed to analyze the automatic evaluation of the recordings of the video game. Three corpus of the recordings of the video game but gathered in different times were used. The first corpus contained recordings of 5 speakers with Down syndrome (605 recordings). The second corpus contained recordings of other 5 speakers with Down syndrome. And the third corpus contained recordings of 13 speakers with Down syndrome obtained with a previous version of the video game. These recordings were evaluated by a prosody expert following the categories of intonational phonology (intonation, accent and prosodic organization). The evaluation was binary, Right (R) or Wrong (W), depending of the quality of the recording. In addition, some recordings were evaluated by the therapist who was next to the players during the game sessions, using a 3-level scoring. If the evaluation was Cont.R (Continue with right result) or Cont. (Continue but the oral activity could be better), the video game advanced to the next activity. If the evaluation was Rep. (Repeat), the game offered a new attempt in which the player had to repeat the activity. With the aim of automatically predicting

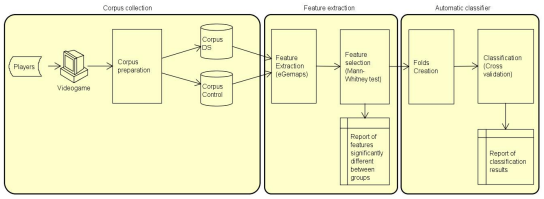


Figure 3: *Scheme of the experimental procedure to compare the speech of people with Down syndrome and the speech of people without intellectual disabilities.*

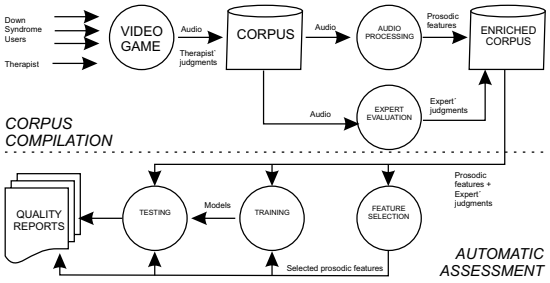


Figure 4: *Experimental procedure scheme followed to analyze the automatic evaluation of the video game recordings.*

the prosody expert evaluations, the same set of features used in the previous experiment (GeMAPS) was extracted from each recording. These features together with the prosody expert evaluations were used to train and test 3 classifiers using 10-folds cross validation. Finally, the automatic and human based evaluations of 5 speakers were compared to analyze the impact of speakers with Down syndrome heterogeneity on classification results.

5. Results and discussion

Related with the design and evaluation of the video game, the results showed that people with Down syndrome had more errors and needed more time to complete the activities (production and comprehension) than people without intellectual disabilities. The data gathered during the game sessions could be useful to detect specific prosodic problems of each player. This data can be used by speech therapists to train the skills related with these activities with the aim of improving them. In addition, players with Down syndrome showed a high motivation to complete the activities of the video game. This is an important result because people with Down syndrome present cognitive limitations that difficult them doing training tasks. The inclusion of the training activities inside a narrative helped to improve the motivation of the players. The questionnaires done by the speech therapists after the game sessions strengthened this results. On the other hand, the video game was useful to gather a speech corpus focused on the prosody of people with Down syndrome.

Table 1 shows the results related with identifying the speaker group (Down syndrome or typical speakers) of each recording of the corpus, using 3 automatic classifiers. The acoustic features related with fundamental frequency, energy and rhythm selected alone show worse performance results than spectral ones, but the combination of these 3 type of features obtained better performance results than the spectral alone. The spectral features are usually related with particularities of the

Table 1: *Classification results for identifying the group of the speaker. Classification rate (C.Rate) and UAR using different feature sets and different classifiers are reported. The features used are those with significant differences between TD and DS groups. The classifiers are decision tree (DT), support vector machine (SVM) and multilayer perceptron (MLP). # is the number of input features in each set.*

		SVM		MLP		DT	
Set	**#**	**C. Rate**	**UAR**	**C. Rate**	**UAR**	**C. Rate**	**UAR**
Frequency	9	62.67	0.61	64.33	0.64	60.17	0.60
Energy	9	79.33	0.78	76	0.76	72.50	0.71
Temporal	9	76.83	0.76	77.83	0.78	74.33	0.75
Frequency+Energy+Temporal	27	90	0.9	91.83	0.91	82	0.82
Spectral	34	87.33	0.87	87.33	0.87	84.33	0.84
All	61	94.17	0.94	95.17	0.95	86.50	0.87

Table 2: *Percentage of coincidence between therapist decision, classifier and prosody expert per speaker. Concerning the classifier, R represents the utterances classified as Right by the classifier and W represents the utterances classified as Wrong by the classifier. Each row percentage is relative to the number of each type of utterances of prosody expert evaluation.*

		Expert judgment		**Classified as**		**Therapist decision**		
Speaker	**#Total utt**	**type**	**#utt**	**R**	**W**	**Cont.R**	**Cont.**	**Rep.**
S01	120	R	87	83.91%	16.09%	68.97%	24.14%	6.90%
		W	33	57.58%	42.42%	30.30%	36.36%	33.33%
S02	106	R	81	87.65%	12.35%	85.19%	14.81%	0.00%
		W	25	28.00%	72.00%	84.00%	16.00%	0.00%
S03	97	R	78	97.44%	2.56%	94.87%	3.85%	1.28%
		W	19	73.68%	26.32%	100.0%	0.00%	0.00%
S04	131	R	75	94.57%	5.33%	21.33%	44.00%	34.67%
		W	56	41.07%	58.93%	5.36%	32.14%	62.5%
S05	151	R	77	87.01%	12.99%	20.78%	50.65%	28.57%
		W	74	29.73%	70.27%	6.76%	20.27%	72.97%
Total	605	R	398	89.96%	10.05%	59.06%	27.14%	13.84%
		W	207	41.06%	58.94%	28.05%	23.72%	48.31%

phonological system of people with Down syndrome, so the training of the use of these features is complicated or impossible. However, the use of the features related with fundamental frequency, energy and rhythm can be improved by therapy. This improving can be reflected on a better use of intonation, accent or rhythm in the speech of people with Down syndrome. In addition, the results of the perceptual test done to evaluate the prosody transferred recordings show the importance of prosody to identify a voice as atypical. The recordings of people without intellectual disabilities with the transferred prosody (frequency, energy and duration) of the recordings of people with Down syndrome were identified mainly as atypical speech as well as the recordings of people with Down syndrome with the transferred prosody of the recordings of people without intellectual disabilities were identified mainly as typical speech.

The results of the experiment to automatically evaluate the prosodic quality of the video game recordings show the impact of the heterogeneity presented by people with Down syndrome in this automatic evaluation (Table 2). In short, agreement between the prosodic expert and the therapist depends on the speaker's developmental levels and the type of sentence produced (right or wrong). In addition, differences in the evaluation context can also explain raters' disagreements. Thus, while the expert only based her decisions on intonational criteria, the therapist also took into consideration the progress of the player while playing the video game. In doing so, avoiding frustration was a priority; therefore, levels of frustration tolerance and number of failures influenced the therapist's decisions. In the video game, it is very important to avoid evaluating as wrong a correct utterance; otherwise, frustration may arise. Bearing in mind that the video game aims to engage and motivate the users, the percentage of false negatives must be as low as possible. These results show that only 10.1% of the samples evaluated as Right by the expert are classified as Wrong by the classifier.

6. Conclusions and future work

Concerning the video game design, the narrative and the training activities were developed in collaboration with experts in intellectual disabilities, experts in language deficits, the therapists of the centers of special education and the final users. This process is important because it allowed the detection of possible errors in the design before the development phase. The evaluation carried out by the people with Down syndrome allowed the extraction of some key aspects for the development of educational tools focused on people with Down syndrome. The realization of the training activities is the main objective when a educational video game is developed, so engagement is a very important usability aspect that has to be taken into account to motivate the players.

In addition, the possibility of recording a speech corpus using the video game is relevant specially when the target population has some cognitive limitation. The corpus gathered allowed the comparison between the prosody of people with Down syndrome and the prosody of people without intellectual disabilities using the acoustic features extracted from the recordings. The experiments carried out using automatic classifiers to identify the speaker group that produced a recording obtained high performance results using features related with fundamental frequency, energy, temporal and spectral domains. On the other hand, the perceptual experiment carried out using the recordings created by the transferred prosody algorithm showed the importance of prosody to identify speech as atypical. This result is important because the prosody training can help people with Down syndrome to improve their integration in society.

Furthermore, the heterogeneity presented by people with Down syndrome affected the automatic evaluation of the recordings quality and the perceptual evaluation done by experts. The concordance values among the therapist, the prosody expert and the automatic classifiers varied depending of the cognitive level and speech quality of each speaker. These results suggest that the automatic classifiers focused on evaluating the prosody quality of people with Down syndrome have to take into account the heterogeneity of this population with the aim of obtaining better performance results.

As future work, the adaptation of the training activities to each speaker profile can improve the learning process. In addition, the generation of a report of the results of each speaker can help therapists to focus on the specific problems of each speaker with the aim of doing a personalized training.

7. Acknowledgements

I want to thank my mentors David Escudero Mancebo and César González Ferreras for their support in the realization of this thesis. This thesis was defended on September 2019. The full paper thesis is available on http://uvadoc.uva.es/handle/10324/38470.

8. References

[1] R. S. Chapman, "Language development in children and adolescents with Down syndrome," *Mental Retardation and Developmental Disabilities Research Reviews*, vol. 3, no. 4, pp. 307–312, 1997.

[2] J. Cleland, S. Wood, W. Hardcastle, J. Wishart, and C. Timmins,

"Relationship between speech, oromotor, language and cognitive abilities in children with Down's syndrome," *International journal of language & communication disorders*, vol. 45, no. 1, pp. 83–95, 2010.

[3] G. E. Martin, J. Klusek, B. Estigarribia, and J. E. Roberts, "Language characteristics of individuals with Down syndrome," *Topics in Language Disorders*, vol. 29, no. 2, p. 112, 2009.

[4] R. D. Kent and H. K. Vorperian, "Speech impairment in Down syndrome: a review," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 1, pp. 178–210, 2013.

[5] V. Stojanovik, "Prosodic deficits in children with Down syndrome," *Journal of Neurolinguistics*, vol. 24, no. 2, pp. 145–155, 2011.

[6] S. J. Peppé, "Why is prosody in speech-language pathology so difficult?" *International Journal of Speech-Language Pathology*, vol. 11, no. 4, pp. 258–271, 2009.

[7] A. McFarlane, A. Sparrowhawk, and Y. Heald, *Report on the educational use of games*. TEEM (Teachers evaluating educational multimedia), Cambridge, 2002.

[8] O. Saz, S. Yin, E. Lleida, R. Rose, C. Vaquero, and W. R. Rodríguez, "Tools and Technologies for Computer-Aided Speech and Language Therapy," *Speech Communication*, vol. 51, no. 10, pp. 948–967, 2009.

[9] C. González-Ferreras, D. Escudero-Mancebo, M. Corrales-Astorgano, L. Aguilar-Cuevas, and V. Flores-Lucas, "Engaging adolescents with Down syndrome in an educational video game," *International Journal of Human–Computer Interaction*, pp. 1–20, 2017.

[10] M. Corrales-Astorgano, D. Escudero-Mancebo, and C. González-Ferreras, "Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome," *Speech Communication*, vol. 99, pp. 90– 100, 2018.

[11] M. Corrales-Astorgano, P. Martínez-Castilla, D. Escudero-Mancebo, L. Aguilar, C. González-Ferreras, and V. Cardeñoso-Payo, "Automatic assessment of prosodic quality in down syndrome: Analysis of the impact of speaker heterogeneity," *Applied Sciences*, vol. 9, no. 7, p. 1440, 2019.

[12] M. T. Lee, J. Thorpe, and J. Verhoeven, "Intonation and phonation in young adults with Down syndrome," *Journal of Voice*, vol. 23, no. 1, pp. 82–87, 2009.

[13] A. Rochet-Capellan and M. Dohen, "Acoustic characterisation of vowel production by young adults with Down syndrome," in *18th International Congress of Phonetic Sciences (ICPhS 2015)*, Glasgow, United Kingdom, Aug. 2015. [Online]. Available: https://hal.univ-grenoble-alpes.fr/hal-01240561

[14] G. Albertini, S. Bonassi, V. Dall'Armi, I. Giachetti, S. Giaquinto, and M. Mignano, "Spectral analysis of the voice in Down syndrome," *Research in developmental disabilities*, vol. 31, no. 5, pp. 995–1001, 2010.

[15] O. Saz, J. Simón, W. Rodríguez, E. Lleida, C. Vaquero *et al.*, "Analysis of acoustic features in speakers with cognitive disorders and speech impairments," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 1, 2009.

[16] K. Bunton and M. Leddy, "An evaluation of articulatory working space area in vowel production of adults with Down syndrome," *Clinical linguistics & phonetics*, vol. 25, no. 4, pp. 321–334, 2011.

[17] J. Van Borsel and A. Vandermeulen, "Cluttering in Down syndrome," *Folia Phoniatrica et Logopaedica*, vol. 60, no. 6, pp. 312–317, 2008.

[18] D. Devenny and W. Silverman, "Speech dysfluency and manual specialization in Down's syndrome," *Journal of Intellectual Disability Research*, vol. 34, no. 3, pp. 253–260, 1990.

[19] K. Eggers and S. Van Eerdenbrugh, "Speech disfluencies in children with down syndrome," *Journal of Communication Disorders*, vol. 71, pp. 72–84, 2018.

[20] L. Zampini, M. Fasolo, M. Spinelli, P. Zanchi, C. Suttora, and N. Salerni, "Prosodic skills in children with Down syndrome and in typically developing children," *International Journal of Language & Communication Disorders*, vol. 51, no. 1, pp. 74–83, 2016.

[21] C. P. Moura, L. M. Cunha, H. Vilarinho, M. J. Cunha, D. Freitas, M. Palha, S. M. Pueschel, and M. Pais-Clemente, "Voice parameters in children with Down syndrome," *Journal of Voice*, vol. 22, no. 1, pp. 34–42, 2008.

[22] M. Moran and H. Gilbert, "Selected acoustic characteristics and listener judgments of the voice of down syndrome adults," *American journal of mental deficiency*, vol. 86, no. 5, p. 553—556, March 1982.

[23] R. Rodger, "Voice quality of children and young people with Down's Syndrome and its impact on listener judgement," Ph.D. dissertation, Queen Margaret University, 2009.

[24] P. Standen, N. Anderton, R. Karsandas, S. Battersby, and D. Brown, "An evaluation of the use of a computer game in improving the choice reaction time of adults with intellectual disabilities," *Journal of Assistive Technologies*, vol. 3, no. 4, pp. 4–11, 2009.

[25] A. Brandão, L. Brandão, G. Nascimento, B. Moreira, C. N. Vasconcelos, and E. Clua, "Jecripe: stimulating cognitive abilities of children with Down syndrome in pre-scholar age using a game approach," in *Proceedings of the 7th International Conference on Advances in Computer Entertainment Technology*. ACM, 2010, pp. 15–18.

[26] P. Standen, F. Rees, and D. Brown, "Effect of playing computer games on decision making in people with intellectual disabilities," *Journal of Assistive Technologies*, vol. 3, no. 2, pp. 4–12, 2009.

[27] M. Cagatay, P. Ege, G. Tokdemir, and N. E. Cagiltay, "A serious game for speech disorder children therapy," in *Health Informatics and Bioinformatics (HIBIT), 2012 7th International Symposium on*. IEEE, 2012, pp. 18–23.

[28] B. Black, "Educational software for children with down syndrome-an update," *Down Syndrome News and Update*, vol. 6, no. 2, pp. 66–68, 2006.

[29] D. Le, K. Licata, C. Persad, and E. M. Provost, "Automatic assessment of speech intelligibility for individuals with aphasia," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 11, pp. 2187–2199, 2016.

[30] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "Peaks–a system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.

[31] A. K. Maier, F. Hönig, C. Hacker, M. Schuster, and E. Nöth, "Automatic evaluation of characteristic speech disorders in children with cleft lip and palate," in *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*. ISCA, 2008, pp. 1757–1760.

[32] I. Laaridh, W. B. Kheder, C. Fredouille, and C. Meunier, "Automatic prediction of speech evaluation metrics for dysarthric speech," in *Proc. Interspeech 2017*, 2017, pp. 1834–1838.

[33] W. Quesenbery, "The five dimensions of usability," *Content and complexity: Information design in technical communication*, pp. 81–102, 2003.

[34] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the Munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[35] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Self-supervised Deep Learning Approaches to Speaker Recognition: A Ph.D. Thesis Overview

Umair Khan and Javier Hernando

TALP Research Center, Department of Signal Theory and Communications,
Universitat Politecnica de Catalunya Barcelona, Spain
{umair.khan, javier.hernando}@upc.edu

Abstract

Recent advances in Deep Learning (DL) for speaker recognition have improved the performance but are constrained to the need of labels for the background data, which is difficult in practice. In i-vector based speaker recognition, cosine (unsupervised) and PLDA (supervised) are the basic scoring techniques, with a big performance gap between the two. In this thesis we tried to fill this gap without using speaker labels in several ways. We applied Restricted Boltzmann Machine (RBM) vectors for the tasks of speaker clustering and tracking in TV broadcast shows. The experiments on AGORA database show that using this approach we gain a relative improvement of 12% and 11% for speaker clustering and tracking tasks, respectively. We also applied DL techniques in order to increase the discriminative power of i-vectors in speaker verification task, for which we have proposed the use of autoencoder in several ways, i.e., (1) as a pre-training for a Deep Neural Network (DNN), (2) as a nearest neighbor autoencoder for i-vectors, (3) as an average pooled nearest neighbor autoencoder. The experiments on VoxCeleb database show that we gain a relative improvement of 21%, 42% and 53%, using the three systems respectively. Finally we also proposed a self-supervised end-to-end speaker verification system. The architecture is based on a Convolutional Neural Network (CNN), trained as a siamese network with multiple branches. From the results we can see that our system shows comparable performance to a supervised baseline.

Index Terms: deep learning, speaker verification, i-vector, autoencoder, CNN, speaker embeddings

1. Introduction

Deep Learning (DL) approaches have shown their success in image and speech technologies which has inspired the community to apply these approaches in speaker recognition as well [1, 2, 3]. The current DL application in speaker recognition can be categorized as: at the frontend, like [4, 5, 6, 7, 8, 9], at the backend, such as in [10, 11, 12], and as an end-to-end system such as in [13, 14, 15]. The most common and the so called speaker embeddings are typically extracted from an intermediate layer of a Deep Neural Network (DNN). The inputs to the network are feature vectors, like the Mel-Frequency Cepstral Coefficients (MFCC) or in some cases spectrograms. Whereas the output of the network is fed with the class (speaker) labels for the background data. Therefore, these DL approaches are typically constrained to labeled background data.

The i-vector representation of speech [16], with cosine scoring, is an unsupervised process. However, Probabilistic Linear Discriminant Analysis (PLDA) [17] is the most efficient backend for i-vectors which leads to a superior performance as compared to cosine scoring but at the cost of labeled background data. However, in practice, it is difficult to access large amount of labeled data. In this thesis [18], we applied self-supervised DL approaches to improve the performance without using speaker labels. We addressed this problem in three different ways.

As a first objective of this thesis, we applied Restricted Boltzmann Machine (RBM) vector representation of speech for the tasks of speaker clustering and tracking in TV broadcast shows. Such a representation is referred to as RBM vector which has shown success in speaker verification task in [19]. In the second objective, we applied self-supervised DL approaches in order to increase the discriminative power of i-vectors for speaker verification. For this purpose we used autoencoder in three different ways, i.e., (1) as a pre-training for a DNN, (2) as a nearest neighbor autoencoder for i-vectors, (3) as an average pooled nearest neighbor autoencoder. In the last main objective of this thesis, we proposed a self-supervised end-to-end speaker verification system. The network architecture is based on a Convolutional Neural Network (CNN) which is trained as a siamese network with multiple branches.

The rest of the paper is organized as follows. Sections 2, 3, and 4 explain the three main objectives of the thesis, respectively. Section 5 describes the experimental setup and results. Section 6 lists the publication resulted from the Ph.D. thesis and section 7 concludes the paper.

2. RBM vectors for speaker clustering and tracking

We have proposed RBMs at the front-end for the tasks of speaker clustering and speaker tracking in TV broadcast shows. RBMs are trained to transform utterances into a vector based representation. Because of the lack of data for a test speaker, we propose RBM adaptation to a global model. First, the speaker independent global model, which is referred to as universal RBM (URBM), is trained with all the available background data. Then a speaker dependent adapted RBM model is trained with the data of each test speaker. The visible to hidden weight matrices of the adapted models are concatenated along with the bias vectors and are whitened using Principal Component Analysis (PCA) to generate the vector representation of speakers. These vectors, referred to as RBM vectors, were shown to preserve speaker-specific information and are used in the tasks of speaker clustering and speaker tracking. Figure 1 shows a visualization of the connection weights of the URBM (top) and of two randomly selected speakers (bottom). From the figure, it is clear that the speaker dependent adapted RBM weights are driven in speaker-specific direction which are discriminative.

For the speaker clustering task, we extract RBM vectors using the method described above for the test speakers. All the speaker segments that are to be clustered are represented by

10.21437/IberSPEECH.2021-38

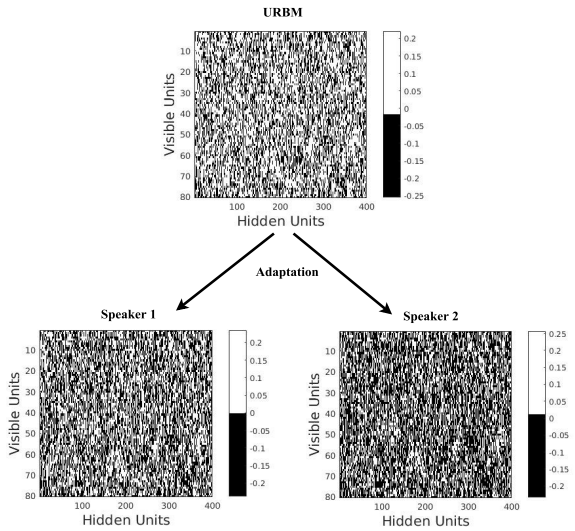


Figure 1: *Comparison of URBM and adapted RBMs weights.*

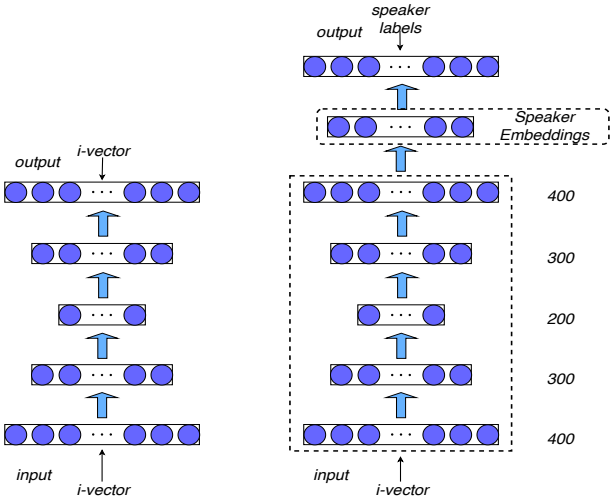


Figure 2: *(left) Autoencoder pre-training (right) DNN training.*

RBM vectors. Then we cluster these RBM vectors by applying a bottom-up Agglomerative Hierarchical Clustering (AHC) approach using cosine and PLDA scores [20]. For the speaker tracking task, we implement a two stage strategy. The first stage is speaker segmentation in which the audio is segmented according to the speaker change points [21]. In the second stage, the segments generated are identified against all the target speakers, in order to specify which segment belongs to which target speaker [21, 22]. We represent all the segments and target speakers by RBM vectors. Then, the RBM vectors of all the segments are scored against the RBM vectors of all the target speakers using cosine and PLDA scoring.

3. Autoencoder based approaches for speaker verification

3.1. Autoencoder as a pre-training for DNN

In this work we have proposed the use of autoencoder pre-training for post-processing of i-vectors in speaker verification task. The conventional architecture of an autoencoder consists of an $encoder$ and a $decoder$ as shown in Figure 2 (left). The $encoder$ is a function that encodes the input i-vector w into a lower dimensional space, and the $decoder$ is a function that decodes it back in order to reconstruct w. In order to avoid the need of large amount of labeled background data, we train the autoencoder using a large amount of unlabeled data. The training is carried out by minimizing the Mean Square Error (MSE) between the input w and the reconstructed $w\hat{}$. Then, we train a DNN classifier using a relatively small labeled data. Therefore, this is a semi-supervised DL approach. We initialize the parameters of the DNN training with the weight matrices and bias vectors of the pre-trained autoencoder. In this way, we train a hybrid autoencoder-DNN classifier. After the training, we transform i-vectors into new representation as the output from the second last layer of the network as shown in Figure 2. The goal is to improve the performance using fewer background speaker labels. The experimental results have shown that the proposed approach has improved the baseline system in two aspects. Firstly, the proposed system outperforms the baseline system in terms of EER. Secondly, we have observed that the hybrid autoencoder-DNN training converges faster as compared to the one without autoencoder pre-training [23].

3.2. Nearest neighbor approach

In this work we have proposed to train an autoencoder to reconstruct neighbor i-vectors instead of the same training i-vectors, as usual. These neighbor i-vectors are selected in an unsupervised manner according to the highest cosine scores to the training i-vectors. In this way the autoencoder learns speaker variability when no labeled background data is available. The conventional training is carried out by minimizing the loss function : $MSE(w\hat{}, w)$. We propose to train the autoencoder by minimizing the loss function : $MSE(w\hat{}, v)$, as shown in Figure 3, where v is a neighbor i-vector of w and $w\hat{} = decoder(encoder(w))$. We propose an automatic selection of the neighbor i-vectors according to the Algorithm 1, as explained in [24].

Once the autoencoder is trained with the selected neighbor i-vectors, we transform the testing i-vectors into a new speaker vector representation. We extract the desired speaker vectors at the output of the autoencoder. These are referred to as autoencoder vectors or shortly ae-vectors. In the experiments, ae-vectors have shown to increase the discriminative quality of i-vectors without using speaker labels.

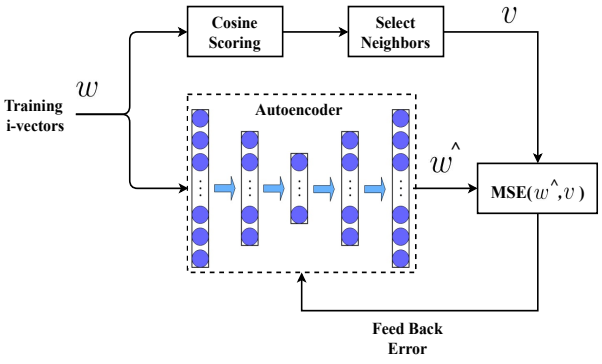


Figure 3: *Proposed training of the autoencoder.*

Algorithm 1: Proposed neighbor i-vectors selection algorithm for a constant k

Input : Training i-vectors w_i, $1 < i < n$
Output: Neighbor i-vectors v_{ij}, $1 < i < n$ and $1 < j < k$

for *each training i-vector* w_i **do**
  for *each training i-vector* w_t, $1 < t < n$ **do**
    if $i \neq t$ **then**
      Compute $score_{i,t} = cosine(w_i, w_t)$
    end
  end
  Select the corresponding k i-vectors with the highest scores as $v_{i,j}$
end

3.3. Average pooled nearest neighbor approach

In this work we train the network using a large set of nearest neighbor i-vectors. For every input i-vector w a set of neighbor i-vectors v_k is input to the network. During the training we minimize the loss between reconstructed $\hat{v}$ and actual training i-vector w. The neighbor i-vectors are selected using the same algorithm but setting a $threshold$ after selecting the constant k number of neighbors. The network architecture is composed of an average pooling layer followed by four fully connected (FC) layers. The input layer is fed by the set of neighbor i-vectors v_k. We train the network by minimizing the loss function $L(\hat{v}, w)$, where $L(\cdot)$ can be Cosine Distance (CD) or MSE, w is the training i-vector and $\hat{v} = f(v_k)$, where $f(\cdot)$ is the non-linearity deployed by the network. During the training, the loss $L(\cdot)$ is back-propagated to the network in every iteration. In this way, the DNN is able to learn from the nearest neighbor i-vectors and avoids using actual speaker labels. After training, we extract speaker vectors for the testing i-vectors, which are used in the experiments [25].

4. End-to-end system

In this work we propose self-supervised siamese networks trained using pairwise training samples, i.e., anchor, client and impostor. Since we do not use speaker labels, we propose to generate the training pairs in an unsupervised manner. The client and impostor selection is carried out in the i-vector space using two databases, i.e., A and B. Suppose Spk_A and Spk_B denote the speakers appearing in database A and B, respectively. We assume that the speakers in database A do not appear in database B, i.e., $Spk_A \cap Spk_B = \phi$.

First, all the i-vectors in A are scored among each other using cosine scoring. For every i-vector in A we select a fix k number of neighbor i-vectors using Algorithm 1 as client i-vectors. After this we apply a $threshold$ to the cosine scores. Then we score all the i-vectors in A with those in B. For every i-vector in A, we select k number of i-vectors from B that are closest according to scores. As the speakers in A do not appear in B, these k selected i-vectors, subjected to a $threshold$, are the impostors i-vectors. In this way, every i-vector in A has been assigned k client and k impostor i-vectors. The network has two identical branches and is trained by minimizing binary cross-entropy loss as shown in Figure 4. The training pairs are in the form [anchor, client] and [anchor, impostor]. After training, we obtain decision scores for the experimental trials at the output of the network. We also trained a triple-branch siamese for which

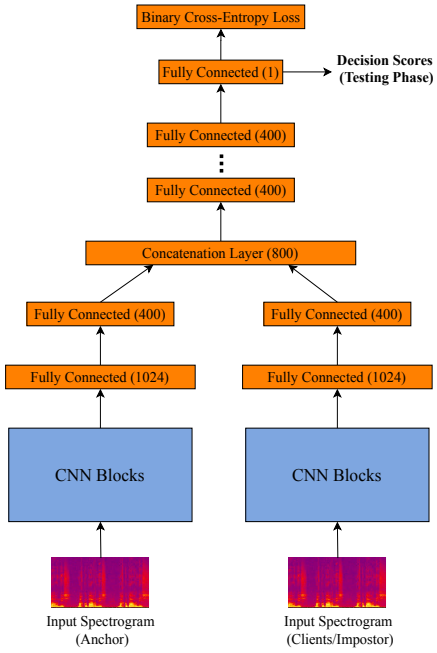


Figure 4: *Block diagram of our end-to-end siamese network.*

we make pairs of three samples, i.e., [anchor, client, impostor], and is trained by minimizing triplet loss. Unlike the double-branch siamese, we extract speaker embeddings from the triple-branch siamese [26].

5. Experimental results

For speaker clustering and tracking experiments we used AGORA Catalan broadcast TV3 dataset [27]. We extracted 2631 speaker segments from the test partition, according to the transcription, for speaker clustering. 414 target speakers were tracked during the tracking experiments. Table 1 shows the results obtained using RBM vectors compared with i-vectors. For speaker clustering an Equal Impurity (EI) of 37.14% is obtained with 2000 dimensional RBM vectors using cosine scoring, gaining a relative improvement of 12% over i-vectors. Similarly an EI of 31.68% is obtained with PLDA scoring which results in a relative improvement of 11% over i-vectors. For speaker tracking Equal Error Rate (EER) of 3.30% and 2.74% were obtained using 2000 dimensional RBM vector using cosine and PLDA scoring respectively, with relative improvements of 11.76% and 7.74% over baseline.

Table 1: *Speaker clustering (EI) and tracking (EER) results.*

Clustering	**EI (Cosine)**	**EI (PLDA)**
[1] i-vector (400)	46.26	36.16
[2] i-vector (800)	42.19	35.91
[3] i-vector (2000)	42.83	35.89
[4] RBM vector (400)	39.66	37.36
[5] RBM vector (800)	40.02	32.36
[6] RBM vector (2000)	37.14	31.68
Tracking	**EER (Cosine)**	**EER (PLDA)**
[7] i-vector (800)	3.74	2.97
[8] RBM vector (2000)	3.30	2.74

Table 2: *Speaker verification results using autoencoder pre-training for DNN, in terms of EER.*

Approach	Scoring	EER(%)
[1] i-vector	Cosine	17.61
[2] i-vector	PLDA	9.54
[3] only-encoder-dnn	Cosine	12.73
[4] conventional-dnn	Cosine	8.58
[5] full-autoencoder-dnn	Cosine	7.51

Table 3: *Speaker verification results using nearest neighbor approach, for different values of k, using cosine scoring.*

Approach	k	EER(%)
[1] i-vector	-	17.61
[2] ae-vector	1	15.32
[3] ae-vector	2	12.36
[4] ae-vector	5	10.62
[5] ae-vector	15	10.20
Fusion of [1] & [5]	-	9.82

For speaker verification we used VoxCeleb-1 [28] dataset for the nearest neighbor and average pooled nearest neighbor approaches, and VoxCeleb-2 [29] dataset for the autoencoder pre-training and end-to-end systems. From the test partition of VoxCeleb-1, 37,720 speaker verification trials were scored for evaluation. Table 2 compares the performance of the autoencoder pre-training with i-vectors. From the table it is clear that our proposed speaker vectors has outperformed the i-vector system by a relative improvement of 21%, in terms of EER. Also, it is shown in [23] that using the pre-training strategy the DNN training convergence was faster than the conventional training. Table 3 shows the results obtained using the nearest neighbor approach for different values of k. From the table we observe that the ae-vectors has gained a relative improvement of 42% over i-vector/cosine. Moreover the EER of 10% is very close to that of i-vector/PLDA (9.54%). Table 4 shows the results obtained using the average pooled nearest neighbor approach for different values of k using CD and MSE losses. We observed that MSE loss is the best choice. Moreover our approach has gained a relative improvement of 53% over i-vector/PLDA at the cost of using background data in testing [25].

Table 5 shows the results obtained using the end-to-end and triple-branch networks. From the Table we can see that as we increase the value of k, the performance improves. The best EER of 6.90% was achieved using k equal to 10. Setting the value of k equal to 10, we have trained our triple-branch siamese network using triplet loss and we extracted speaker embeddings. The triple-branch siamese network has achieved an EER of 6.95%. A score level fusion gives an EER of 6.07% which is very close to the supervised AMSoftmax baseline [26, 30].

6. Publications

[1] Umair Khan, Pooyan Safari, and Javier Hernando. Restricted Boltzmann Machine Vectors for Speaker Clustering. In *Proc. IberSPEECH*, pages 10–14, 2018, **(Awarded the ISCA travel grant).**

[2] Umair Khan, Pooyan Safari, and Javier Hernando. Restricted boltzmann machine vectors for speaker clustering and tracking tasks in tv broadcast shows. *Applied Sciences*, 9(13):2761, 2019.

Table 4: *Speaker verification results using average pooled nearest neighbor approach with CD and MSE, using cosine scoring.*

k	CD Loss	MSE Loss
[1] 10	8.81	8.70
[2] 20	6.60	6.56
[3] 30	5.68	5.64
[4] 50	4.97	4.98
[5] 100	4.84	4.45
[6] 150	6.53	4.48

Table 5: *Speaker verification results using end-to-end and triple-branch siamese, in comparison to supervised baselines.*

Approach	k	EER(%)
[1] i-vector/PLDA	-	9.54
[2] Baseline (AMSoftmax)	-	5.71
[3] End-to-end	2	7.81
[4] End-to-end	5	7.73
[5] End-to-end	10	6.90
[6] Triple-branch	10	6.95
Fusion of [5] & [6]	10	6.07

[3] Umair Khan and Javier Hernando. Dnn speaker embeddings using autoencoder pre-training. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.

[4] Umair Khan, Miquel India, and Javier Hernando. Autoencoding nearest neighbor i-vectors for speaker verification. *Proc. Interspeech 2019*, pages 4060–4064, 2019.

[5] Umair Khan, Miquel India, and Javier Hernando. i-vector transformation using k-nearest neighbors for speaker verification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7574–7578. IEEE, 2020.

[6] Umair Khan and Javier Hernando. Unsupervised training of siamese networks for speaker verification. *Proc. Interspeech 2020*, pages 3002–3006, 2020.

[7] Umair Khan and Javier Hernando. The upc speaker verification system submitted to voxceleb speaker recognition challenge 2020 (voxsrc-20). *arXiv preprint arXiv:2010.10937*, 2020, **(3rd prize winner of self-supervised track).**

7. Conclusions

The contributions of this thesis are presented in three main objectives. Firstly, the use of RBM vectors for speaker clustering and tracking, which resulted in a relative improvement (RI) of 12% and 11%, respectively. Secondly, we applied DL techniques to improve i-vectors for speaker verification, in several ways. The experimental results show that we gain a RI of 21%, 42% and 53%. Finally we trained an end-to-end speaker verification system, which showed a comparable performance to supervised baseline. From the thesis we conclude that using DL approaches, despite of being unsupervised, we could do better in scenarios where labels are not available for the training data.

8. Acknowledgements

This work was supported by the project PID2019-107579RB-I00 / AEI / 10.13039/501100011033

9. References

[1] K. Chen and A. Salman, "Learning speaker-specific characteristics with a deep neural architecture," *IEEE Transactions on Neural Networks*, vol. 22, no. 11, pp. 1744–1756, 11 2011.

[2] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 10 2015.

[3] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Proc. Odyssey*, 2014, pp. 293–298.

[4] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, "Deep feature for text-dependent speaker verification," *Speech Communication*, vol. 73, pp. 1–13, 2015.

[5] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[6] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[7] G. Bhattacharya, J. Alam, and P. Kenny, "Deep speaker embeddings for short-duration speaker verification," *Proc. Interspeech 2017*, pp. 1517–1521, 2017.

[8] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *Proc. Interspeech 2018*, pp. 2252–2256, 2018.

[9] S. Novoselov, A. Shulipa, I. Kremnev, A. Kozlov, and V. Shchemelinin, "On deep speaker embeddings for text-independent speaker recognition," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 378–385.

[10] M. Senoussaoui, N. Dehak, P. Kenny, R. Dehak, and P. Dumouchel, "First attempt of boltzmann machines for speaker verification," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[11] T. Stafylakis, P. Kenny, M. Senoussaoui, and P. Dumouchel, "Plda using gaussian restricted boltzmann machines with application to speaker verification," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[12] O. Ghahabi and J. Hernando, "Deep learning backend for single and multisession i-vector speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 807–817, 4 2017.

[13] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5115–5119.

[14] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances." in *Interspeech*, 2017, pp. 1487–1491.

[15] S. Dey, S. R. Madikeri, and P. Motlicek, "End-to-end text-dependent speaker verification using novel distance measures." in *Interspeech*, 2018, pp. 3598–3602.

[16] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[17] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[18] U. Khan, "Self-supervised deep learning approaches to speaker recognition," Ph.D. dissertation, Universitat Politècnica de Catalunya, 2021.

[19] P. Safari, O. Ghahabi, and J. Hernando, "From features to speaker vectors by means of restricted boltzmann machine adaptation," in *ODYSSEY 2016-The Speaker and Language Recognition Workshop*, 2016, pp. 366–371.

[20] U. Khan, P. Safari, and J. Hernando, "Restricted Boltzmann Machine Vectors for Speaker Clustering," in *Proc. IberSPEECH*, 2018, pp. 10–14.

[21] U. Khan, "Speaker tracking system using speaker boundary detection," Master's thesis, Universitat Politècnica de Catalunya, 2016.

[22] U. Khan, P. Safari, and J. Hernando, "Restricted boltzmann machine vectors for speaker clustering and tracking tasks in tv broadcast shows," *Applied Sciences*, vol. 9, no. 13, p. 2761, 2019.

[23] U. Khan and J. Hernando, "Dnn speaker embeddings using autoencoder pre-training," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[24] U. Khan, M. India, and J. Hernando, "Auto-encoding nearest neighbor i-vectors for speaker verification," *Proc. Interspeech 2019*, pp. 4060–4064, 2019.

[25] U. Khan, M. India, and J. Hernando, "I-vector transformation using k-nearest neighbors for speaker verification," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7574–7578.

[26] U. Khan and J. Hernando, "Unsupervised training of siamese networks for speaker verification," *Proc. Interspeech 2020*, pp. 3002–3006, 2020.

[27] H. Schulz and J. A. R. Fonollosa, "A catalan broadcast conversational speech database," in *Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages*, 2009, pp. 27–30.

[28] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Interspeech*, 2017.

[29] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech*, 2018.

[30] U. Khan and J. Hernando, "The upc speaker verification system submitted to voxceleb speaker recognition challenge 2020 (voxsrc-20)," *arXiv preprint arXiv:2010.10937*, 2020.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

A study of data augmentation for increased ASR robustness against packet losses

María Pilar Fernández-Gallego[1], *Doroteo T. Toledano*[1]

[1]AUDIAS - Audio, Data Intelligence and Speech
Universidad Autónoma de Madrid

mariapilar.fernandezg@estudiante.uam.es, doroteo.torre@uam.es

Abstract

Nowadays a large amount of companies record conversations, calls, sales or even meetings, in many cases to comply with the current legislation. Apart from the legal need, these recordings constitute an invaluable source of information about clients, call center operators, marketing campaigns, markets trends, etc. The current state of the art in Automatic Speech Recognition (ASR) allows to exploit this information in a very efficient way. However, the recordings at these repositories tend to present very low quality because the audio is typically recorded in a highly compressed way to save storing space. Besides, since it is very common to use Voice over IP (VoIP) in these systems, it is usual to have short interruptions in the speech signal due to packet losses. Both effects, and particularly the last one, have an impact in ASR performance.

This paper presents an extensive study of the influence of these effects and the effectiveness of different data augmentation strategies to increase the robustness of ASR systems in these circumstances, and in particular when packet losses degrade the speech signal.

Index Terms: Data augmentation, Packet losses, Speech recognition, Fisher Spanish

1. Introduction

Nowadays many companies record conversations, calls, sales or even meetings for several reasons and, in many cases, just to comply with the current legislation. Apart from the legal need, these recordings constitute an invaluable source of information about clients, call center operators, marketing campaigns, markets trends, etc. Typically the speech recorded in call centers is recorded in a very reduced bit rate, and hence limited quality. The main reason for it is to save storing space because in some cases hundreds or thousands of hours daily need to be recorded. Therefore reduced bit rate speech coding is a common scenario in Automatic Speech Recognition (ASR) applied to this type of data.

Voice over IP (VoIP) communications have become very common and are currently mainstream in call centers and voice recording systems. In VoIP, speech signals are transmitted as packets of a fixed length that depends on the selected speech codec. Normally the packet length is between 20 and 40 ms. During the transmission of these packets several issues can occur, the most common being packet delays and packet losses, which may degrade the speech quality beyond the degradation introduced by speech coding and other common problems such as echoes [1]. To mitigate losses or delay issues, Packet Loss Concealment (PLC) techniques are applied, which can fill the lost packets by adding redundant information such as repeating frames, adding noise, etc; or alternatively using interpolation methods to reconstruct the the signal [2] [3]. PLC techniques are included in some standards, such as ITU-G711 Appendix I, which is a high quality low-complexity algorithm for packet loss concealment [4].

In the last years, deep learning techniques and deep neural networks (DNNs) have proved to be able to extract higher level features from less processed data and also to model, predict and generalize better than classical techniques. In ASR in particular, DNNs are included in all state-of-the-art systems, either combined with the Hidden Markov Model (HMM) machinery in hybrid HMM-DNN systems or completely replacing HMMs with end-to-end neural approaches. One of the reasons for the popularity of DNNs in ASR is the fact that, when training data is abundant and representative of the application, they are very robust against variability, such as bandwidth, environment or speaker [5]. However, to attain these advantages it is necessary to have a large amount of training data. For that reason, data augmentation methods have become commonly used to artificially generate more data, trying to represent the possible scenarios that a system may face in real operation. In a good number of works, data augmentation has been used to increase model robustness against variations such as noise, speed, reverberation, etc., with significant improvements with respect to other techniques [6][7]. However there are very few works in which it is applied to deal with the problem of packet losses and reduced bit rate codification. Some works like [8] try to use data augmentation to increase robustness to deformations in the time direction and partial loss of frequency information by working on log mel spectrogram directly. Other works such as [9] apply data augmentation using audio codecs with changed bit rate, sampling rate, and bit depth.

Therefore, the motivation for this study on data augmentation to increase ASR robustness against packet losses and reduced bit rate coding is based on two main reasons. The first one is the need to develop a robust ASR system capable of dealing with packet losses and limited bit rate speech codification, which are very common in call center scenarios, without a huge degradation in performance. The second one is that after studying the available scientific literature, we noticed that the research community has not extensively addressed this problem, despite it is a very common issue in the industry.

The rest of the paper is organized as follows: Section 2 explains the types of data augmentation used. Section 3 describes the ASR system used and data sets used to train it. Section 4 describes the data used to carry on the experiments and the results obtained. Finally, Section 5 presents conclusions and future work.

2. Augmentation types

To make ASR models more robust against packet losses and low quality coding we need to artificially introduce these dis-

10.21437/IberSPEECH.2021-39

tortions in the training database for the ASR acoustic models. This section describes how these distortions are introduced in the training data.

2.1. Packets losses

To simulate packet losses we have considered a fixed frame size of 20 ms. We consider three modes for the packet losses: individual packet losses, burst packet losses and mixed (individual and burst) packet losses. A tunable parameter controls the percentage of packets lost, which in our experiments take the values 5%,10%,15% or 20%.

2.1.1. Individual packet losses

To simulate this type of packet loss, randomly chosen packets along the audio file are removed (by making the signal 0), assuring that the removed packets are not together.

2.1.2. Burst packet losses

To recreate burst packet losses we remove batches of three consecutive frames randomly located along the audio until we reach the loss percentage chosen.

2.1.3. Single and burst packet losses

A real scenario can include both types of packet losses. For this reason, a more realistic simulation has been considered merging the two previous modes (single and burst losses), by removing randomly located batches of one, two or three packets along the audio until the loss percentage selected is reached.

2.2. Speech coding

To simulate the effect of speech coding we have used the two more common codecs in our real data: MP3 (with FFmpeg [10]) and full rate GSM (with SoX [11]), with several variations.

2.2.1. MP3

MP3 is a perceptual audio (not speech specific) codec designed for audio transmission and storage that became very popular for Internet audio applications and streaming in particular. Its encoding efficiency is defined by the bit rate, which can be adjusted to the particular needs of the application among several possible choices [12]. In this paper, two bit rates have been applied: 16 Kbit/s and 8 Kbit/s. To obtain 8 Kbit/s files, each channel of the original audio has been to converted to 8Kbit/s MP3 format and then to WAV-PCM. For the 16 kbit/s files, each channel has been converted to 16 kbit/s MP3 and later converted to WAV-PCM. In both cases FFmpeg [10] has been used.

2.2.2. Full rate GSM 06.10

GSM FR is a speech codec designed for digital mobile telephone use. The speech signal is divided into blocks of 20 ms and has an average bitrate of 13 kbps using a 8 kHz sampling rate [13]. Each channel is converted to GSM FR format and then coverted to WAV-PCM format using SoX [11].

3. Automatic Speech Recognition Models

KALDI [14] has been used for training the ASR models. In particular, the acoustic model is a Hybrid Deep Neural Network Hidden Markov Model (DNN-HMM) which uses Time Delay Neural Networks (TDNN) [15]. This type of network includes connections between its units not only at the current time, but also at different times, and is capable of taking non-linear decisions taking into account the value of the input at a relatively long time span, normally around the current time. Each layer works with a time context wider than the previous layer thus using an increasingly amount of temporal context. For language modelling we have used a relatively simple n-gram statistical language model. Our recipe is based on the Fisher/Callhome Spanish recipe included in Kaldi without applying the re-scoring last stage.

The architecture of the TDNN consists of 13 TDNN layers with 1024 units each and 128 bottleneck dimensions, a pre-final linear layer with 192 dimensions and a last softmax layer. Input features are Mel frequency cepstral coefficients (MFCCs) with high frequency resolution (hires) with 40 dimensions, adding a ± 1 temporal context and an i-vector with 100 dimensions to model speaker characteristics.

In addition, the recipe applies speed-perturbation for data augmentation with two factors 0.9 and 1.1 obtaining a data set three times bigger than the original (including the original and two speed-perturbed copies). [6]

3.1. Baseline

The data used to develop our baseline system has been the Fisher Spanish data set. This data set is made up with 163 hours of telephone speech from 136 native Caribbean and non-Caribbean Spanish speakers. Around 4 hours have been reserved for the test set and another 4 hours for development. The rest of the corpus has been used for training. The baseline includes speed-perturbation data augmentation. The language model has been trained only on the transcriptions of this corpus.

3.2. Data augmentation experiments

We have retrained the system using different data augmentation strategies to deal with the problems of low bit rate coding and packet losses. In all of them, the amount of data used to train the models are exactly twice the one used for the baseline model. We refer to the different data augmentation strategies explored as da_model1-da_model4.

3.2.1. Low bit rate codec

This strategy (da_model1) trains the system using the baseline data plus another transformed Fisher Spanish data set obtained by applying each of the three codecs described in Section 2.2 to one third of the original corpus. The goal of this data augmentation strategy is to measure the gain obtained by including low bit rate coding effects.

3.2.2. Packets losses and low bit rate coding

To deal with both problems, these strategies double the training data by generating and additional copy in which each file has randomly suffered one or two transformations, low bit rate coding and/or packet losses of different types, depending on the model of packet losses applied:

- da_model2: Individual packet losses (Section 2.1.1).
- da_model3: Burst packet losses (Section 2.1.2).
- da_model4: Individual and burst packet losses (Section 2.1.3).

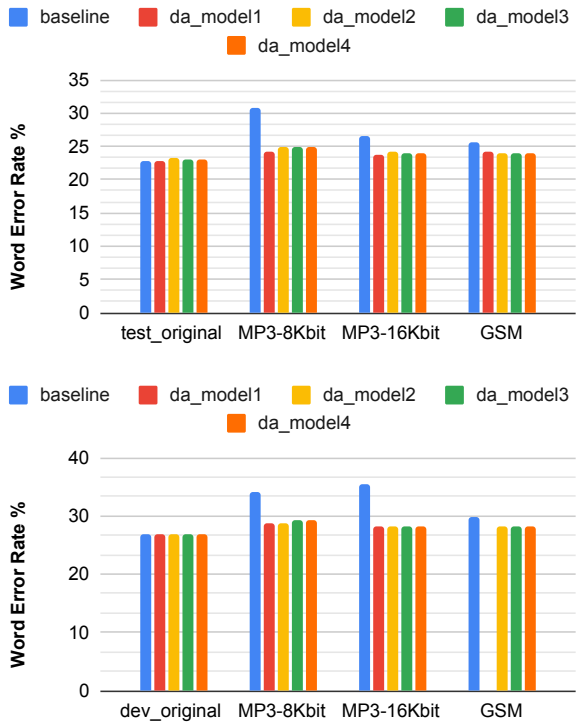
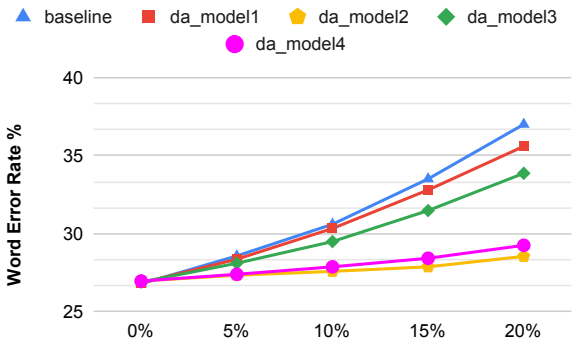
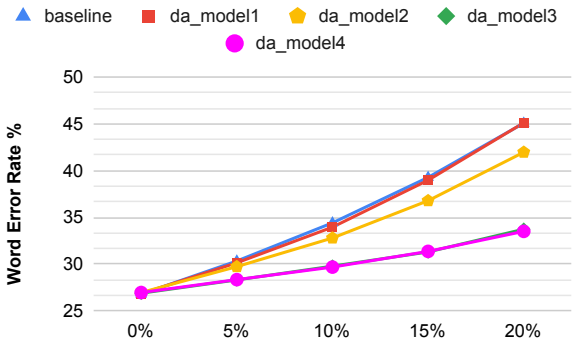


Figure 1: *Baseline vs. different data augmentation systems in test (top) and development (bottom) sets in original format and with three different low bit rate codecs (MP3-8Kbit/s, MP3-16KBit/s and GSM-FR).*



(a) *Single packets loss*



(b) *Burst packets loss*

Figure 2: *Development results comparing the baseline with data augmentation systems for single packet losses (up) and burst packet losses (bottom) for different packet loss probabilities. da_model3 is hidden behind da_model4 in bottom figure.*

4. Experiments and Results

The experiments performed try to measure the effectiveness of the different data augmentation approaches to deal with the problems of packet losses and low bit rate coding. We aim to improve the Word Error Rate (WER) of ASR both on simulated data and also on real data coming from real call centers.

4.1. Evaluation database

The main evaluation database is derived from the test and development datasets reserved from Fisher Spanish. These datasets have been transformed with the same transformations used in the different data augmentation strategies described above.

The three different codecs explained in Section 2.2, have been used to obtain three copies coded with MP3 with 8 and 16 Kbit/s and with GSM.

The different packet loss simulation systems have been used to obtain eight different copies of the data sets in each case applying 5%, 10%, 15% and 20% loss percentage including individual and burst packet losses, so that we can evaluate the worse scenario when the audio file only contains burst packet losses and a better scenario where there are only single packet losses.

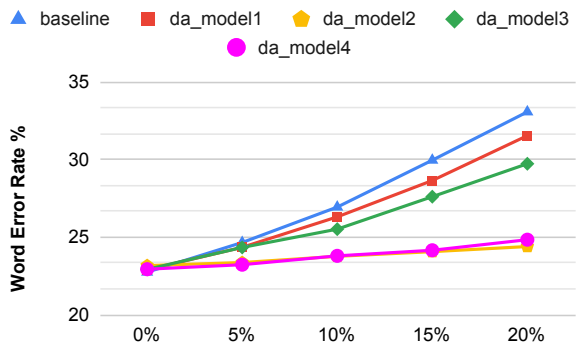
As a result, 12 conditions with different degradations are obtained. Table 1 shows the Mean Opinion Score (MOS) (as estimated by ITU-T P.563 single-ended method for objective speech quality assessment in narrow-band telephony applications [16]) in each one of this conditions, and also for the non-degraded baseline. The MOS scale is from 1 to 5, being the worst score 1 and the best 5. It is somewhat surprising that single packet loss data get worse MOS than burst packet loss data, probably because the percentage is the same but there are three times more interruptions for the individual packet losses.

4.2. Low bit rate coding results

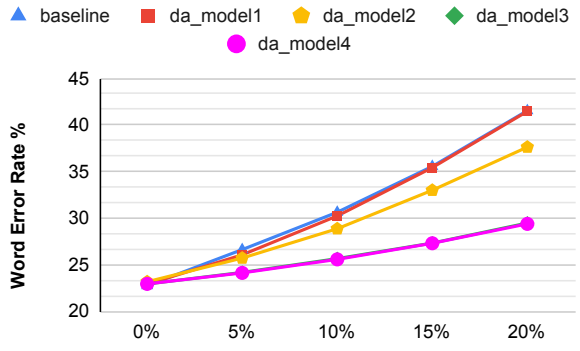
Figure 1 shows a big difference in terms of WER for low bit rate coding data sets, particularly between the baseline and all the data augmentation models, the latter having small differences among them. A much better WER is obtained when data augmentation techniques are applied, achieving almost the same results obtained with the baseline system in the non-degraded scenario. There is a small difference of $\sim 1\%$ in all cases, except when coding with MP3 at 8 Kbit/s where the difference is $\sim 2\%$. Therefore, it seems that these types of low bit rate coding scenarios can be almost solved with any of these data augmentation techniques.

4.3. Packet loss results

Figures 2 and 3 show the results obtained with the baseline and the different systems trained with different data augmentation strategies in development and test data sets, respectively, both for individual packet losses and burst packet losses at different packet loss probabilities. The data augmentation strategy including single losses only (da_model2) has better results in scenarios with only individual losses. Similarly, the data augmentation strategy containing burst losses (da_model3) has bet-



(a) *Single packets loss*



(b) *Burst packets loss*

Figure 3: *Test results comparing the baseline with data augmentation systems for single packet losses (up) and burst packet losses (bottom) for different packet loss probabilities. da_model3 is hidden behind da_model4 in bottom figure.*

ter results in the scenario with burst losses, showing in this case a great improvement compared with the rest of the models. Finally, the data augmentation strategy with single and burst losses (da_model4) have similar results in both cases than the best of the former models, and therefore shows an improved robustness against all types of packet losses.

When packet loss probability is relatively low (5%, 10%) and only individual losses are considered, ASR can be very accurate using data augmentation techniques, having similar results as without packet losses. However, when the number of packet losses increases or burst packet losses occur, data augmentation techniques greatly improve performance, but still a notable decrease in performance is observed. In the worst-case scenarios, performance could be about $\sim 7\%$ worse in terms of absolute WER than with non-degraded data, even for the best performing data augmentation strategies.

So far all the experiments were performed on simulated data including simulated degradations. With the aim to corroborate the effectiveness of these data augmentation strategies in real data, we have applied the baseline (no data augmentation besides the speed perturbation) and the best performing data augmentation strategy of the previous experiments, the da_model4, in real call-center data. Language model and lexicon have been adapted to the particularities of the specific call center data. In all cases the language model and lexicon used in the baseline and the da_model4 are exactly the same. Table 2 shows the results obtained on real call center data. Three data sets coming

Table 1: *Mean Opinion Score (MOS). PL means packet loss*

Transformation	**Test**	**Dev**
Original	2, 27	2, 42
MP3 8 Kbit/s	2, 03	2, 19
MP3 16 Kbit/s	1, 95	1, 87
GSM	1, 85	1, 95
5% Indiv. PL	1, 27	1, 43
10% Indiv. PL	1, 14	1, 24
15% Indiv. PL	1, 06	1, 11
20% Indiv. PL	1, 03	1, 04
5% Burst PL	1, 40	1, 61
10% Burst PL	1, 19	1, 33
15% Burst PL	1, 10	1, 19
20% Burst PL	1, 06	1, 10

from three different call centers have been used to compare ASR performance using the baseline and the data augmentation strategy. In all of them, results are clearly better with the da_model4 model, even 10% in the test_call_center_3 data set compared to the baseline model. This indicates that these strategies are not only adequate to model simulated low bit rate coding and packet losses problems, but also to improve recognition performance in real data.

Table 2: *Results in real data.*

Dataset	**duration (h)**	**baseline**	**da_model4**
		WER %	
test_call_center_1	3.5	39, 51	32, 07
test_call_center_2	3	40, 68	34, 20
test_call_center_3	2	47, 83	37, 94

5. Conclusions

In this work, we have applied data augmentation techniques to mitigate reduced ASR accuracy in audio including low bit rate coding and packet losses, having found that data augmentation by itself can greatly improve robustness in these scenarios, which are very common in the industry.

Experiments have shown that data augmentation can be very effective when audios include low bit rate codecs or contain relatively infrequent individual packet losses, obtaining an important improvement compared to the baseline model and reaching results close to those obtained without these degradations. However, for more frequent packet losses or burst losses, degradations in terms of WER remain important despite data augmentation mitigation, making it advisable to research other alternatives to try to compensate even more these degradations.

As future work, we plan to study techniques to recover loss packets to try to attain better accuracy in ASR when there are very frequent packet losses, including burst losses.

6. Acknowledgements

Partly funded by project RTI2018-098091-BI00, Ministry of Science, Innovation and Universities (Spain) and FEDER.

7. References

[1] A. Takahashi, H. Yoshino, and N. Kitawaki, "Perceptual QoS assessment technologies for VoIP," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 42, no. 7, pp. 28–34, 2004.

[2] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE network*, vol. 12, no. 5, pp. 40–48, 1998.

[3] B. W. Wah, X. Su, and D. Lin, "A survey of error-concealment schemes for real-time audio and video transmissions over the Internet," in *Proceedings - International Symposium on Multimedia Software Engineering*, 2000.

[4] "Recommendation, I. T. U. T. G. 711, Appendix I, a high quality low-complexity algorithm for packet loss concealment with G. 711," *Int. Telecom. Union (ITU)*, 1999.

[5] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition," in *Proceedings of International Conference on Learning Representation*, 2013.

[6] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition." in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6-10, Dresden, Germany, Proceedings*, 2015, pp. 3586–3589.

[7] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[8] D. S. Park, W. Chan, Y. Zhang, B. Z. Chung-Cheng Chiu, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *INTERSPEECH 2019 – The 20th Annual Conference of the International Speech Communication Association, Graz, Austria, Sep. 15-19, Proceedings*, 2019, pp. 2613–2617.

[9] N. Hailu, I. Siegert, and A. Nürnberger, "Improving automatic speech recognition utilizing audio-codecs for data augmentation," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020, pp. 1–5.

[10] S. Tomar, "Converting video formats with FFmpeg," *Linux Journal*, vol. 2006, no. 146, p. 10, 2006.

[11] C. Bagwell. (1996) SoX - Sound eXchange. [Online]. Available: http://sox.sourceforge.net/ (Accessed: 4 February 2021)

[12] P. Noll, "MPEG digital audio coding," *IEEE Signal Processing Magazine*, vol. 14, no. 5, pp. 59–81, 1997.

[13] "ETSI EN 300 961 Digital cellular telecommunications system (Phase 2+)(GSM); Full rate speech; Transcoding, GSM 06.10 version 8.1. 1 Release 1999." in *Proceedings of International Conference on Learning Representation*, vol. 8, no. 1, 2000.

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, and M. Hannemann, "The KALDI speech recognition toolkit," *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[15] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6-10, Dresden, Germany, Proceedings*, 2015, pp. 3214–3218.

[16] L. Malfait, J. Berger, and M. Kastner, "P. 563—The ITU-T standard for single-ended speech quality assessment." *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

TRIBUS: An end-to-end automatic speech recognition system for European Portuguese

Carlos Carvalho[1,2], *Alberto Abad*[1,2]

[1]INESC-ID, Lisbon, Portugal
[2]Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
carlos.f.carvalho@tecnico.ulisboa.pt, alberto.abad@inesc-id.pt

Abstract

End-to-end automatic speech recognition (ASR) approaches have emerged as a competitive alternative to traditional HMM-based ASR systems. Unfortunately, most end-to-end ASR systems are not easily reproduced since they require vast amounts of data and computational resources that are only available for a reduced set of companies and labs worldwide. Consequently, the performance of these systems is not very well known for low resource languages to the best of our knowledge. European Portuguese is one of those languages. In this work, we present a set of experiments to train and assess some of the current most successful end-to-end ASR approaches for European Portuguese. The proposed system, named TRIBUS, is a hybrid CTC-attention end-to-end ASR combining data from three different domains: read speech, broadcast news and telephone speech. For comparison purposes, we also train a state-of-the-art HMM-based baseline on the same data. Experimental results show that TRIBUS achieves 8.40% character error rate (CER) on the broadcast news test set without the need of a language model, in contrast to the 4.33% CER attained by the HMM baseline on the same set using an in-domain language model. We consider this result quite promising, especially for highly unpredictable vocabulary ASR applications.

Index Terms: automatic speech recognition, end-to-end, hybrid CTC-attention, low resources

1. Introduction

Speech recognition technology is submerged in our society more than ever. Products like Siri, Cortana, Google Now and Amazon Echo Alexa which belong to big companies, like Apple, Microsoft, Google and Amazon, respectively, are part of our every day lives. This high tech translates into a significant number of applications (e.g., healthcare [1] and autonomous vehicles [2]) which have contributed to increase the quality of live in our society.

Traditionally, large vocabulary continuous speech recognition (LVCSR) systems, i.e., HMM-based systems, rely on sophisticated modules including acoustic, phonetic and language models, which are manually created by specialized computational linguists and engineers . Since all these modules do not optimize the same goal, the ASR system final objective typically has more difficulties in achieving a global optimum. Furthermore, HMM systems and n-gram language models make conditional independence assumptions, whereas real speech does not follow those strict assumptions. To overcome these limitations, it is possible to replace the HMM-based system with a single deep neural network, which is trained following a global optimization procedure. Also, by removing the engineering required for the usual alignment, bootstrapping, clustering and decoding with finite-state transducers (FSTs), characteristic of most HMM-based systems, the training and decoding process becomes more straightforward. This new paradigm, named end-to-end, directly maps an input sequence of acoustic features to an output sequence of tokens, i.e., characters or sub-words, [3, 4, 5, 6].

Some widely used contemporary end-to-end approaches are: connectionist temporal classification (CTC) [3, 7], attention encoder-decoder (AED) [5, 8] and RNN Transducer (RNN-T) [9]. CTC's main problem is that it is not capable of modelling language [10] because it considers each label in the output sequence to be independent of each other. To solve CTC-based models independence assumption, the RNN-T approach was proposed [9]. In contrast to CTC, RNN-T does not make assumptions about label independence when enumerating the hard alignments. However, the main disadvantage of CTC-based and RNN-T systems is that, since they first enumerate all hard alignments and then aggregate them, there could be many illogical paths. Attention-based models solve this problem by creating a direct soft alignment between input and output, with the support of an attention mechanism. One of the main issues of attention-based models is the monotonic alignment problem. As a result, the attention mechanism can allow extremely nonsequential alignments between input frames and output tokens [11]. To solve this, hybrid CTC-attention models were proposed in [12]. These models use the advantages of both CTC-based and attention-based architectures in training and decoding.

The main drawback of these end-to-end systems, mentioned above, is that they require a considerable number of training hours to achieve state-of-the-art performance results when compared to traditional HMM-based systems [4]. For English ASR, corpora such as TED-LIUM [13], and Librispeech [14] offer great possibilities for researchers to experiment and compare large end-to-end ASR systems. However, this is not the case for European Portuguese (EP), mainly due to the lack of large scale speech data resources publicly available, either paid or for free.

The main contribution of this work is the development and assessment of the first known end-to-end ASR system for EP in a low resource scenario, by using one of the most successful end-to-end ASR approaches. All corpora used for the experiments of this work correspond to small to medium sized data sets collected by INESC-ID over the past years. For comparison purposes, we also report results obtained with a conventional HMM system trained on the same data.

The remainder of the paper is organized as follows. We start by describing the corpus used to train the end-to-end sys-

This work has been partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020. Authors would like to thank Pedro Esteves and Thomas Rolland.

10.21437/IberSPEECH.2021-40

tem and the HMM-based baseline in Section 2. Section 3 gives a brief description of the acoustic feature extraction and a description of the central architecture used to train the end-to-end ASR TRIBUS system. Section 4 details the experimental setup including the baseline system, the results and a comparison between the proposed model and baseline. Finally, a concluding summary is presented in Section 5.

2. EP corpus

This section provides a detailed overview of the speech data resources collected by INESC-ID that helped to create the TRIBUS corpus, followed by a description of the language and pronunciation models used.

2.1. Speech data resources and partitions

The TRIBUS corpus training set is a collection of three training sets from three datasets representing different domains: read speech, broadcast news and telephone read speech. INESC-ID participated in the design, collection, processing, transcription and/or distribution of these corpora over the past years. The validation and test sets of the TRIBUS corpus used in the present work are the original ones from each corpus, except for telephone read speech data, where the design process will be detailed below:

Read speech: The read speech corpus used is BD-PÚBLICO [15]. Similar to Wall Street Journal corpus [16], BD-PÚBLICO was created from the Portuguese newspaper Público in 1997. BD-PÚBLICO contains 120 different speakers, where ages range from 19 to 28 years old. It is particularly designed for research and development of speaker-independent continuous speech recognition approaches. The training set contains around 23 hours with 100 speakers and 8389 utterances, and the validation and test set contains 2 hours and 10 speakers each. The validation set contains 584 utterances and the test set contains 592 utterances. Finally, the sampling rate of this corpus is 16kHz.

Broadcast news speech: The EP broadcast news (BN) corpus used in this work is ALERT [17], which contains spontaneous speech from BN shows. ALERT was created in cooperation with RTP, a public service broadcasting organization from Portugal. The training data contains around 60 hours of speech with 1366 speakers and 47552 utterances. The validation set contains 8 hours of data with 260 speakers and 6222 utterances, and the test set has 6 hours has 175 speakers and 4701 utterances. Finally, this corpus is sampled at 16kHz.

Telephone speech: The telephone speech data considered belongs to the SPEECHDAT corpus [18], a collection of read speech utterances from telephone calls, collected by Portugal Telecom, a Portuguese telecommunications operator currently known as Altice Portugal. SPEECHDAT contains two main recording phases: SPEECHDAT 0 and SPEECHDAT 1. Each telephone call included in the database contains 33 read items and 7 spontaneous answers, where some contain demographic information. Only 9 phonetically rich sentences, from the set of 33 items, were used in this work. As opposite to ALERT and BD-PÚBLICO, an experimental setup for SPEECHDAT was created. When working with SPEECHDAT, we noticed that from all the 36243 utterances from SPEECHDAT 1 only 3622 are unique, and from the total 9000 utterances from SPEECHDAT 0 only 904 are unique. Furthermore, SPEECHDAT 0 and SPEECHDAT 1 are two disjoint sets. For this reason, SPEECHDAT 1 was chosen for the training set, and we divided SPEECHDAT 0 into two parts: the validation set and the test set. This data splitting process was made such that the number of female and male speakers was approximately the same for each set. Overall, the training set contains approximately 63 hours with 4027 speakers and 36243 utterances, and the validation and test set contains 9 hours each. Finally, the sampling rate of this corpus is 8kHz.

Table 1: *Summary of the number of utterances, speakers and hours for the training, validation and test set of the TRIBUS corpus.*

	#utterances	**#speakers**	**#hours**
Training set	92184	5493	146
Validation sets			
ALERT	6222	260	8
BD-PÚBLICO	584	10	2
SPEECHDAT	4497	500	9
Test sets			
ALERT	4701	175	6
BD-PÚBLICO	592	10	2
SPEECHDAT	4503	501	9

The total amount of hours, number of utterances and number of speakers in the training, validation and test partitions of the TRIBUS corpus are presented in Table 1. All data was downsampled to 8kHz to match the telephone data sampling rate.

2.2. Language and pronunciation models

The language model (LM) for each set, used in the HMM-based baseline, is the one that comes with each corpus, except for SPEECHDAT. ALERT language model was designed by interpolating three distinct LMs. The first is a backoff 4-gram LM, trained on a word corpus of newspapers texts containing 700M words. This out-of-domain corpus was collected from the web. The second LM is a backoff 3-gram LM trained on an in-domain corpus of broadcast news transcripts, with around 531k words. Finally, the third model is a backoff 4-gram LM, estimated from the EP web newspapers, collected the week before creating the interpolated LM. The final interpolated LM is a 4-gram LM with Kneser-Ney modified smoothing, 100k 1-gram, 7.5M 2-gram, 14M 3-gram and 7.9M 4-gram. Following, BD-PÚBLICO language model is a backoff 3-gram closed model.

To create the language model for SPEECHDAT, we first estimated a backoff 3-gram model with Kneser-Ney smoothing combined with Good-Turing smoothing. To avoid over-fitting due to the small linguistic variability in the training set, mentioned above, we interpolated this 3-gram LM model with BD-PÚBLICO LM [15]. An additional step was performed to normalize the notation of all the noise (e.g., _nsnoise_) and disfluencies (e.g., _ehm_hmm_) across the three datasets. Finally, for the TRIBUS corpus, we collected a lexicon of 108358 pronunciations, obtained from publicly available resources.

3. End-to-end model for EP ASR

First, we will describe how the acoustic features are created. Next, the main idea behind the attention architecture used will be mentioned, and finally, the end-to-end hybrid CTC-attention system named TRIBUS, depicted in Figure 1, will be described.

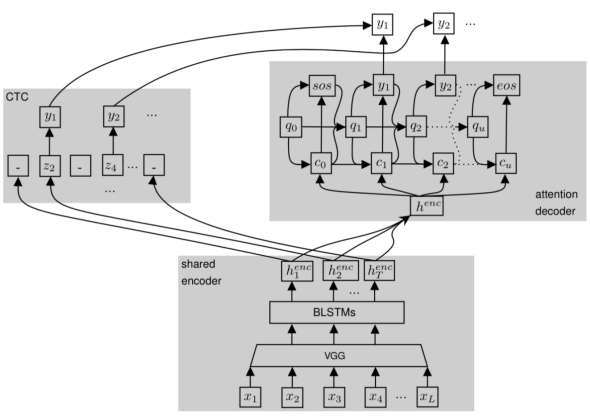


Figure 1: *TRIBUS hybrid CTC-attention architecture. Adapted from [22].*

3.1. Acoustic features

The acoustic features consist of 80-dimensional Mel filterbank energies plus 3 additional pitch features, extracted with Kaldi [19], making the final size of the acoustic vector equal to 83.

3.2. Attention-based architecture

The attention architecture contains three models: the encoder, the hybrid attention mechanism and decoder. The encoder network is described as follows:

$$\boldsymbol{h}_t^{enc} = \textit{Encoder}(\boldsymbol{x}), \tag{1}$$

which is in charge of converting the input features $\boldsymbol{x}$ into a framewise hidden vector $\boldsymbol{h}_t^{enc}$. Then, the hybrid attention weight is computed as:

$$\alpha_{ut} = \textit{Hybrid attention}(\boldsymbol{q}_{u-1}, \{\alpha_{u-1}\}_{t=1}^{T}, \boldsymbol{h}_t^{enc}), \tag{2}$$

where α_{ut} is the weight that says how much attention is going to vector $\boldsymbol{h}_t^{enc}$, in order to compute output y_u, and $\boldsymbol{q}_{u-1}$ is the last hidden state of the long short-term memory (LSTM) [20] present in the decoder network, mentioned with more detail below. After computing all weights corresponding to all framewise hidden vectors $\boldsymbol{h}_t^{enc}$, we compute a weighted summation of hidden vectors $\boldsymbol{h}_t^{enc}$ to form the hidden vector $\boldsymbol{c}_u$:

$$\boldsymbol{c}_u = \sum_{t=1}^{T} \alpha_{ut} \boldsymbol{h}_t^{enc}. \tag{3}$$

At last, the decoder uses the weighted summation $\boldsymbol{c}_u$ and the last output y_{u-1} to compute the new output y_u:

$$p(y_u|y_1 ... y_{u-1}, \boldsymbol{x}) = \textit{Decoder}(\boldsymbol{c}_u, y_{u-1}). \tag{4}$$

3.2.1. Encoder network

The encoder network used, Eq. 1, consists of two initial blocks of the VGG layer [21]. With this, the number of frames is reduced approximately by a factor of 4. Following, there are 4 BLSTM layers with 1024 hidden and output units. Each BLSTM layer is followed by a linear projection layer and the final output is 1024 features for every reduced frame.

3.2.2. Hybrid attention mechanism

The hybrid attention mechanism in Eq. 2 is decomposed as:

$$\{\boldsymbol{f}_t\}_{t=1}^{T} = \boldsymbol{K} * \boldsymbol{\alpha}_{u-1}, \tag{5}$$

where each $\boldsymbol{f}_t$ is a vector of size 10 and $*$ denotes a 1D convolution operation along axis t, with the convolution parameter $\boldsymbol{K}$, to produce the set of features $\{\boldsymbol{f}_t\}_{t=1}^{T}$.

Then, we can compute the energy value as:

$$\begin{aligned} e_{ut} = \boldsymbol{g}^T tanh(&\text{LinearNN}(\boldsymbol{q}_{u-1}) \\ &+ \text{LinearNNB}(\boldsymbol{h}_t^{enc}) \\ &+ \text{LinearNN}(\boldsymbol{f}_t)), \end{aligned} \tag{6}$$

where LinearNN corresponds to a linear transformation with learnable matrix parameters and LinearNNB to an affine transformation, with learnable matrix and bias vector parameters. The number of output features used for the three linear networks is 320. Then, we use all e_{ut} values and apply a softmax function to get the attention weight α_{ut}, so that we can compute the target output y_u:

$$\alpha_{ut} = Softmax(\{e_{ut}\}_{t=1}^{T}). \tag{7}$$

3.2.3. Decoder network

The decoder network (Eq. 4) computes:

$$Decoder(.) = Softmax(\text{LinearNNB}(LSTM_u)). \tag{8}$$

The $LSTM_u$ is conditioned on three variables. The first is the previous hidden state $\boldsymbol{q}_{u-1}$. The second is the ground truth character, y_{u-1}, which is extracted from an embedding layer, trained while training the full end-to-end network. At last, the third is the attention vector $\boldsymbol{c}_u$, which is concatenated with the previous character vector, giving a vector of size 2048 as input to the $LSTM_u$ cell.

For this architecture, two LSTM cells with 1024 units were used, where the new hidden state $\boldsymbol{q}_u$ is computed as:

$$\boldsymbol{q}_u = LSTM_u(\boldsymbol{q}_{u-1}, \boldsymbol{c}_u, y_{u-1}). \tag{9}$$

3.3. Hybrid CTC-attention network

In the hybrid CTC-attention architecture, the CTC and attention decoder networks share the same encoder. Also, when training, the CTC and attention loss are combined, to achieve more robustness and converge faster [12]:

$$Loss_{Total} = \lambda Loss_{CTC} + (1 - \lambda) Loss_{Attention}, \tag{10}$$

where $\lambda \in [0, 1]$. The λ value used for all end-to-end experiments is 0.2, the same as in [23].

3.4. Additional details

All noise and disfluencies from the TRIBUS corpus mentioned in Section 2 are mapped to a special token named <noise>. Also, it is important to note that there are special tokens for CTC and attention-based systems among all other output characters that exist, respectively. CTC requires a <blank> token [7], and the attention architectures requires the *start-of-sentence* and *end-of-sentence* (<sos/eos>) token. Therefore, the full hybrid CTC-attention system will have two special tokens plus an unknown token, <unk>, to map out-of-vocabulary (OOV) symbols. Finally, the total number of output symbols for TRIBUS is 49.

4. Experiments

For all end-to-end experiments, we used the second version of ESPnet toolkit [22] to implement and investigate our proposed methods, which is still under development by the time we write this paper. To evaluate the performance of our end-to-end ASR TRIBUS system, we compared it to a robust HMM-DNN baseline using the same corpus trained with Kaldi [19].

4.1. End-to-end experiments

The model was trained for at most 30 epochs, with early stopping (patience of 4 epochs) based on the validation loss. The training process takes approximately 9 hours in a single GeForce GTX 1080 Ti. Adadelta [24], an adaptive learning rate back-propagation algorithm, was the optimizer chosen, with an initial learning rate of 1.0, a mini-batch size of 30 and gradient clipping of 5. All weights were initialized using Xavier initialization [25]. It was also used a scheduler for the learning rate, where the scale factor was 0.5 and the patience 1 epoch. For data augmentation, we used speed perturbed factors of 0.9, 1.0 and 1.1, and SpecAugment [26]. For SpecAugment, F and T are set to 20 and 100 respectively, m_F and m_T are both set to 2, and W is set to 5. The decoding process of the hybrid CTC-attention model follows the setup in [22]. It is relevant to note that no language model was used when decoding with the end-to-end ASR system.

The word error rate (WER) results for the end-to-end TRIBUS ASR system, are presented in Table 2. From the results, we can see that BD-PÚBLICO has the lowest WER, mainly because it is read speech.

Table 2: *WERs [%] and CERs [%] on the end-to-end and HMM-based ASR systems, using TRIBUS corpus.*

	valid (WER)	test (WER)	test (CER)
HMM-GMM			
ALERT	33.26	34.89	-
BD-PÚBLICO	10.21	11.78	-
SPEECHDAT	8.42	13.49	-
HMM-DNN			
ALERT	9.69	9.65	4.33
BD-PÚBLICO	2.56	3.04	0.95
SPEECHDAT	2.49	4.86	3.26
End-to-end			
ALERT	18.80	19.40	8.40
BD-PÚBLICO	8.60	9.10	2.70
SPEECHDAT	21.20	20.00	8.40

4.2. HMM-based experiments

To create a robust HMM-based baseline for the TRIBUS corpus, we designed a similar procedure to the 's5' recipe of WSJ corpus, from Kaldi. First, to create the alignments for the HMM-DNN system, we trained an HMM-GMM system using the TRIBUS corpus, mentioned in Section 2. The training stages that created the HMM-GMM system are the following: (1) monophone stage, (2) triphones + delta + delta-delta stage, (3) triphones + LDA + MLLT stage and finally, (4) the triphones + SAT stage. For the first training stage (1), only the 2000 shortest utterances from the training set were used. For the second (2), a subset of 30000 utterances from the total of 92184, mentioned in Section 2, were used. Finally, for the last two training stages, (3) and (4), all utterances were used. The results for the last HMM-GMM system trained are depicted in Table 2. After creating the HMM-GMM system, the HMM-DNN system was trained following the Chain recipe from WSJ ("run_tdnn_1i.sh"), in Kaldi. The main difference is that only 12 layers were used to train the model, instead of 13 layers.

The WER results for the HMM-DNN ASR system, with respect to the TRIBUS corpus, are presented in Table 2. From results, we can notice that BD-PÚBLICO achieves the lowest WER. When comparing with the WERs from the end-to-end ASR system, we can observe that there is still a significant difference between the traditional HMM-DNN ASR systems and the end-to-end ASR systems for this low resourced scenario. In contrast to the end-to-end system, these results are obtained using the matched in-domain LMs described in section 2.2. In fact, the performance difference is more noticeable in the telephone data, for which the limited linguistic variability allows the LM to have a strongest impact. Finally, it is worth noticing that the baseline HMM system outperforms by a large margin the last reported result for ALERT [17], where we were able to decrease the absolute WER from 23.50% to 9.65%.

4.3. HMM-based vs end-to-end EP systems

For a better comparison between the end-to-end ASR and the HMM-DNN ASR systems, we also report results using CER in Table 2. We notice that the relative difference in terms of CERs of the proposed end-to-end system with respect to the HMM baseline are similar for the broadcast and read speech domain, but significantly better for the telephone speech domain. For instance, for ALERT, the performance degradation both in terms of WER and CER is approximately a factor of two, while for SPEECHDAT there is a factor of around 2.5 and 4 in terms of CER and WER, respectively. As pointed out previously, the LM seems to have a stronger impact in the telephone domain. Nevertheless, we can conclude that the HMM-based systems are still better than the end-to-end systems in this low resource setting.

5. Conclusions

In this paper, we presented the first known work using state-of-the-art end-to-end ASR systems for low resource EP. From the experimental results, and as it could be expected from the literature, we observe that the TRIBUS end-to-end performance is still far from the HMM-DNN system performance in such a low-resource setting. HMM-DNN systems have the advantage of using an in-domain language model that comprises almost all linguistic variation present in validation and test sets, and a pronunciation dictionary as well. Nonetheless, it is quite remarkable how the end-to-end hybrid CTC-attention systems can learn so much using less than 150 hours of training data and without any language model or pronunciation dictionary.

Overall, end-to-end ASR systems are known to have difficulties to generalize when in the presence of novel data [4], limiting its potential applicability to low resource scenarios. Thus, this problem needs to be specifically addressed in the future, for instance, based on a new architecture with better priors or, perhaps, considering new unsupervised learning algorithms able to create better representations.

6. References

[1] C. Pérez, Y. Campos-Roca, L. Naranjo, and J. Martín, "Diagnosis and tracking of parkinson's disease by using automatically extracted acoustic features," *J Alzheimers Dis Parkinsonism*, vol. 6, no. 260, pp. 2161–0460, 2016.

[2] J. Ivaneckỳ and S. Mehlhase, "An in-car speech recognition system for disabled drivers," in *International Conference on Text, Speech and Dialogue*. Springer, 2012, pp. 505–512.

[3] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *PMLR*, 2014, pp. 1764–1772.

[4] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[5] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016, pp. 4960–4964.

[6] Z. Xiao, Z. Ou, W. Chu, and H. Lin, "Hybrid ctc-attention based end-to-end speech recognition using subword units," in *ISCSLP*, 2018, pp. 146–150.

[7] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.

[8] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *NIPS*, vol. 1, no. 9, pp. 577–585, 2015.

[9] A. Graves, "Sequence transduction with recurrent neural networks," *CoRR*, vol. abs/1211.3711, 2012.

[10] A. Y. Hannun, A. L. Maas, D. Jurafsky, and A. Y. Ng, "First-pass large vocabulary continuous speech recognition using bi-directional recurrent dnns," *arXiv preprint arXiv:1408.2873*, 2014.

[11] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: First results," *arXiv preprint arXiv:1412.1602*, 2014.

[12] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[13] A. Rousseau, P. Deléglise, and Y. Estève, "Ted-lium: an automatic speech recognition dedicated corpus." in *LREC*, 2012, pp. 125–129.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *ICASSP*, pp. 5206–5210, 2015.

[15] J. P. Neto, C. A. Martins, H. Meinedo, and L. B. Almeida, "The design of a large vocabulary speech corpus for portuguese," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[16] D. B. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman*, 1992, pp. 899–902.

[17] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "Audimus. media: a broadcast news speech recognition system for the european portuguese language," in *International Workshop on Computational Processing of the Portuguese Language*, 2003, pp. 9–17.

[18] H. Hoge, H. S. Tropf, R. Winski, H. van den Heuvel, R. Haeb-Umbach, and K. Choukri, "European speech databases for telephone applications," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, 1997, pp. 1771–1774.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE Signal Processing Society*, 2011.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[22] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *Proc. Interspeech*, pp. 2207–2211, 2018.

[23] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[24] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[25] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[26] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

mintzai-ST: Corpus and Baselines for Basque-Spanish Speech Translation

Thierry Etchegoyhen, Haritz Arzelus, Harritxu Gete Ugarte,
Aitor Alvarez, Ander González-Docasal, Edson Benites Fernandez

Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
Mikeletegi Pasalekua, 57, Donostia/San Sebastián
{tetchegoyhen,harzelus,hgete,aalvarez,agonzalezd,eebenites}@vicomtech.org

Abstract

The lack of resources to train end-to-end Speech Translation models hinders research and development in the field. Although recent efforts have been made to prepare additional corpora suitable for the task, few resources are currently available and for a limited number of language pairs. In this work, we describe mintzai-ST, a parallel speech-text corpus for Basque-Spanish in both translation directions, prepared from the sessions of the Basque Parliament and shared for research purposes. This language pair features challenging phenomena for automated speech translation, such as marked differences in morphology and word order, and the mintzai-ST corpus may thus serve as a valuable resource to measure progress in the field. We also describe and evaluate several ST model variants, including cascaded neural components, for speech recognition, machine translation, and end-to-end speech-to-text translation. The evaluation results demonstrate the usefulness of the shared corpus as an additional ST resource and contribute to determining the respective benefits and limitations of current alternative approaches to Speech Translation.
Index Terms: Speech Translation, Basque, Spanish, Corpus

1. Introduction

Deep Learning approaches to natural language processing have achieved significant results in the past years, notably in the fields of machine translation [1, 2] and speech recognition [3, 4]. The possibility to train end-to-end models, in particular, has proved fruitful, building on the ability of artificial neural networks to jointly learn different aspects of a task under a data-driven approach.

Speech translation (ST) has been traditionally modelled under a cascaded approach, chaining automated speech recognition (ASR) and machine translation (MT) systems, with task-specific to optimise information communication between components [5, 6, 7]. Although the dominant approach to the task, systems relying on cascaded components are prone to error propagation and end-to-end neural approaches to ST have been explored in recent years as an alternative [8, 9, 10]. While preliminary results with end-to-end systems have shown promises, cascaded systems still obtain better results overall on standard evaluation datasets [11]. One of the main factors for this state of affairs is the scarcity of parallel speech-text and speech-speech corpora, in contrast with the comparatively larger amounts of data available to train separate ASR and MT models, for some language pairs at least.

The paucity of resources to train end-to-end ST systems has led to recent efforts in developing additional parallel corpora suitable for the task, notably the multilingual MuST-C [12] and Europarl-ST [13] corpora. For some language pairs, however, no parallel resources are readily available, as is the case for Basque-Spanish, for instance.[1] This is particularly limiting considering that languages such as Basque exhibit a number of marked linguistic properties, notably rich morphology and relatively free word order, which can represent a challenge for natural language processing tasks in general, and automated translation in particular [15, 14].

In this work, we describe mintzai-ST, a Basque-Spanish parallel speech-text corpus based on the plenary sessions of the Basque Parliament and shared with the scientific community. We evaluate its usefulness for speech translation, by training and evaluating state-of-the art models under both cascaded and end-to-end approaches. To our knowledge, these are the first comprehensive results in terms of speech translation for this language pair, made possible by the production of specific datasets for the task. With approximately 480 hours of audio and 3.3M target words for Spanish to Basque, 190 hours and 1.5M words for Basque to Spanish, the mintzai-ST corpus enables end-to-end training and comparisons with cascaded models, thus providing a basis for further research on alternative approaches to speech translation in a relatively difficult language pair.

The remainder of this paper is organised as follows: Section 2 describes the data acquisition process and statistics of the corpus; in Section 3 we describe the different models that were built for Basque-Spanish speech translation, including cascaded and end-to-end models; Section 4 discusses the results of the comparative model evaluation on the mintzai-ST datasets; finally, Section 5 draws the main conclusions from this work.

2. Corpus description

The corpus was created from the proceedings of the Basque Parliament from 2011 to 2018, a period of time where audio, transcriptions and translations were publicly available. Speakers expressed themselves in either Basque or Spanish, with a majority of interventions in the latter overall; professional transcriptions and translations were then produced into Spanish and Basque, respectively. We describe below the main processes related to data acquisition and preparation of the final corpus.

2.1. Data acquisition

Raw data were first obtained by crawling the web sites where the official plenary sessions are made available: transcriptions and translations were available at http://www.legebiltzarra.eus; videos of the sessions at: https://www.irekia.euskadi.eus.

Texts from the sessions were available as bilingual PDF files, with content in each language provided in a dedicated col-

[1] The only potential speech translation resource we are aware of for this language pair is the EuskoParl corpus [14], which is based on similar sources as mintzai-ST with the main differences that it is not a parallel speech-text corpus and, to our knowledge, is not publicly available.

10.21437/IberSPEECH.2021-41

umn: one for the transcription of the session and the other for its translation. The content was extracted from the PDF files with PDFtoText[2] and boilerplate removal was performed with in-house content-specific scripts. Since the translations were made at the paragraph level for the most part [14], paragraph-level information was maintained.

Videos were provided in different formats over the years (.flv, .webm and .mp4), and audio extraction was performed with FFmpeg[3]. The mapping between videos and reports was performed via inferences from the respective files metadata whenever possible. In most cases, the available information was not sufficient to map video and PDF files with absolute confidence, with multiple and sometimes duplicate videos in many cases[4]; manual revision and mapping were therefore performed throughout this task.

The statistics for the collected raw data are shown in Table 1.

Table 1: *mintzai-ST: raw data statistics*

YEAR	VIDEOS	PDF	HOURS	WORDS
2011	43	21	86.51	132,595
2012	38	21	117.94	173,199
2013	67	38	215.00	306,621
2014	60	30	176.83	252,887
2015	41	27	134.10	195,112
2016	38	21	113.85	170,608
2017	49	33	173.57	250,862
2018	34	26	128.38	207,910
TOTAL	370	217	1,146.18	18,625,252

2.2. Alignment and filtering

As a first step, metadata were filtered from the PDF-extracted text, and source and target files were extracted from the text in the original columns, preserving paragraph-level alignments. Speaker information was usually located at the beginning of a paragraph and was extracted when available.[5]

As a second step, language identification was performed on each paragraph. Since any error at this stage would propagate to subsequent processes, special attention was paid to ensuring correct language identification by employing two separate tools on the content: TextCat and the language identifier of the OpenNER project [16].[6] Paragraphs were discarded if either tool produced different results as their topmost identified language, or if neither tool identified either one of the expected languages; in all other cases, we retained the identified language, by either one or both of the tools.

The third step involved forced alignment, where each word in the source transcription was aligned to a section of the corresponding audio file via source and time indications. This step was performed with the Kaldi toolkit [17], using a bilingual model to reduce the impact of remaining language identification uncertainties. Alignment was performed with different beam sizes (10, 100, 1000 and 10000) and all content was aligned.

At this stage, the source and target files were split on the basis of the previous alignment information, with one paragraph per file. Forced alignment was then applied again, this time with a monolingual model and a small beam size of 1, with a retry beam of 2, to discard alignment errors and non-literal transcriptions.

Since translation models require specific sentence-based training bitexts, the previously aligned paragraphs were further prepared with sentence splitting, tokenisation and truecasing. All operations were performed with the scripts from the Moses toolkit [18]. Sentence-level alignments were then computed with the Hunalign toolkit [19], with an alignment probability of 0.4.

The filtered and aligned data were then randomly split into train, dev and test subsets of triplets consisting of audio, transcription and translation. Triplets were removed from the test sets if the transcription-translation pair appeared in the training set as well. This measure was adopted to account for the fact that even minor acoustic differences might make a triple differ from another, even though the transcription-translation pair would be a duplicate for the machine translation component. Stricter removal along the previous lines allowed for a fair comparison between cascaded and end-to-end models, and made for a more difficult test set as it mainly discarded acoustic variants of greetings and salutations.

The statistics for the mintzai-ST corpus, for Basque to Spanish and Spanish to Basque, are shown in Table 2. The corpus is shared under the Creative Commons CC BY-NC-ND 4.0 license and is available at the following address: https://github.com/vicomtech.[7]

3. Speech translation models

In this section, we describe the different models and components built to measure the usefulness of the prepared corpus as a basis for speech translation, on the one hand, and to identify the relative benefits and limitations of the different modelling alternatives on the other hand. Two main approaches are described in the next section: cascaded models, based on state-of-the art components for speech recognition and machine translation, and end-to-end neural speech translation models.

3.1. Cascaded models

The speech-to-text cascaded models are based on separate components for speech recognition and machine translation, each trained on their own datasets, either on the in-domain mintzai-ST corpus only or on a combination of the corpus with additional data. We describe each component in turn below.

For the additional dataset, we favoured publicly available corpora close to the mintzai-ST domain which would allow for a straightforward reproduction of our results. For this language pair, only text-based datasets were available with these characteristics and we selected the OpenDataEuskadi corpus [15], prepared from public translation memories by the translation services of the Basque public administration.[8] This corpus is close enough to the mintzai-ST domain to be meaningfully combined and large enough to contribute significantly to differ-

[2]This specific extraction tool was selected as it preserved column-based alignments.

[3]https://www.ffmpeg.org/

[4]For each session, there were between 1 and 7 videos, and 1 and 2 PDF files.

[5]182 speakers were identified overall.

[6]Available at the following addresses, respectively: https://github.com/Trey314159/TextCat and https://github.com/opener-project/language-identifier.

[7]The providers of the original content have granted permission for its use without additional restrictions.

[8]The corpus is available at the following address: http://hltshare.fbk.eu/IWSLT2018/OpendataBasqueSpanish.tgz

Table 2: *mintzai-ST: final corpus statistics*

SRC	TGT	PARTITION	HOURS	SENTENCES	SRC WORDS	TGT WORDS
ES	EU	TRAIN	468.16	175,826	4,512,294	3,328,172
EU	ES	TRAIN	180.96	85,409	1,149,803	1,536,695
ES	EU	DEV	2.60	1,000	25,359	18,566
EU	ES	DEV	2.23	1,000	13,831	18,673
ES	EU	TEST	7.89	2,788	74,758	55,283
EU	ES	TEST	6.35	2,300	37,706	51,003

ent components of the cascaded models. The corpus amounts to 963,391 parallel sentences, with 23,413,116 words in Spanish and 17,802,212 in Basque.

To connect the components, the best hypothesis of the ASR model was fed to the MT model, after generating punctuation as described in Section 3.1.1. Although considering alternative hypotheses in the n-best ASR output might provide additional robustness and accuracy to the overall system, we left an evaluation along these lines for future research.

3.1.1. Speech recognition

Two speech recognition architectures, based on end-to-end models and Kaldi based systems, were trained and evaluated to test their performance on the new compiled corpus.

The end-to-end speech recognition systems were based on the Deep Speech 2 architecture [20] for both languages, and were set up with a sequence of 2 layers of 2D convolutional neural networks followed by 5 layers of bidirectional GRU layers and a fully-connected final layer. The output corresponds to a *softmax* function which computes a probability distribution over characters. Stochastic Gradient Descent (SGD) was used as optimiser and the input consisted of Mel-scale based spectrograms, which were dynamically augmented through the SpecAugment technique [21]. The models were estimated using only audios lasting less than 40 seconds, due to training memory constraints, with a learning rate of 0.0001 annealed by a constant factor of 1.08 for a total of 60 training epochs.

The Kaldi recognition systems were built with the *nnet3* DNN setup, and using the so-called *chain* acoustic model based on a factorised time-delay neural network (TDNN-F) [22], which reduces the number of parameters of the network by factorising the weight matrix of each TDNN layer into the product of two low-rank matrices. Our TDNN-F models consisted of 16 TDNN-F layers with an internal cell-dimension of 1536, a bottleneck-dimension of 160 and a dropout schedule of '0,0@0.2,0.5@0.5,0'. The number of training epochs was set to 4, with a learning rate of 0.00015 and a minibatch size of 64. The input vector corresponded to a concatenation of 40 dimensional high-resolution MFCC coefficients, augmented through speed (using factors of 0.9, 1.0, and 1.1) [23] and volume (with a random factor between 0.125 and 2) [24] perturbation techniques, and the appended 100 dimensional iVectors.

Language models were 5-gram models with modified Kneser-Ney smoothing, estimated with the KenLM toolkit [25], and were used as either a component of Kaldi-based systems or to rescore the end-to-end models' hypotheses.

Finally, the capitalisation models were trained with the recasing tools provided by the Moses open-source toolkit [18], while the punctuation module consisted of a bidirectional RNN model built with out-of-domain data from the broadcast news domain and using the Punctuator2 toolkit [26].

3.1.2. Machine translation

All machine translation models in the experiments reported below were based on the Transformer architecture [2], built with the MarianNMT toolkit [27].

The models consisted of 6-layer encoders and decoders and 8 attention heads. The Adam optimiser was used with parameters $\alpha = 0.0003$, $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. The learning rate was set to increase linearly for the first 16,000 training steps and decrease afterwards proportionally to the inverse square root of the corresponding step. The working memory was set to 8000MB and the largest mini-batch was automatically chosen for a given sentence length that fit the specified memory. The validation data were evaluated every 5,000 steps for models trained on larger out-of-domain datasets and every epoch otherwise; training ended if there was no improvement in perplexity after 5 consecutive checkpoints. Embeddings were of dimension 512, tied between source and target, and all datasets were segmented with BPE [28], using 30,000 operations.

3.2. End-to-end models

End-to-end ST models were trained on the in-domain speech-text corpus, using the Fairseq-ST toolkit[9], which supports different types of sequence-to-sequence neural models [29]. The variant selected for the experiments is the Transformer model enhanced for ST described in [30], more specifically the variant the authors refer to as the S-Transformer.

The model follows the standard Transformer architecture with 6 layers self-attentional encoder and decoder, but adds layers prior to the Transformer encoder, to model 2D dependencies. The audio input is provided to the model in the form of sequences of MEL filters, encoded first by two CNNs to model 2D-invariant features, followed by two 2D self-attention layers to model long-range context. The output of the stacked 2D self-attention layers undergoes a linear transformation, followed by a ReLU non-linearity, and is summed with the positional encoding, prior to feeding the Transformer encoder.

We diverged from the implementation described in [30] on one important aspect. Character-based decoding was replaced with subword decoding, using the previously described BPE models, as the former faced consistent issues, resulting in subpar performance; an identical setup with subwords produced significantly better results overall. Further exploration of these differences between translation pairs is left for future research.

4. Comparative evaluation

We first performed an evaluation centred on cascaded models, where a number of variants could be prepared based on different ASR approaches or different types and volumes of training data.

[9] https://github.com/mattiadg/FBK-Fairseq-S

The variants include: ASR models trained with either an end-to-end neural model (E2E) or the Kaldi toolkit (KAL); ASR and MT models trained on either in-domain data only (IND) or on a combination of in-domain and out-of-domain data (ALL), by integrating the OpenDataEuskadi dataset to train the language and casing models for speech recognition, and the translation models for the MT component; MT models obtained by fine-tuning (FT) models trained on the out-of-domain dataset with the in-domain data.

The results for the cascaded variants on the mintzai-ST test sets, in terms of word error rate (WER) and BLEU [31], are shown in Table 3. All results in the table were computed with ASR output that includes punctuation, generated with the previously mentioned punctuation models. To measure the impact of punctuation on the overall process, differences between BLEU scores obtained with and without punctuation, in that order, are also shown in the table (ΔPUNC).

Table 3: *Evaluation results on cascaded variants*

LANG	ASR	MT	WER	BLEU	ΔPUNC
EU-ES	E2E IND	IND	14.43	28.4	+1.0
EU-ES	E2E ALL	IND	14.12	28.4	+0.8
EU-ES	E2E IND	ALL	14.43	33.3	+2.4
EU-ES	E2E ALL	ALL	14.12	33.4	+2.3
EU-ES	E2E IND	FT	14.43	33.3	+2.4
EU-ES	E2E ALL	FT	14.12	33.4	+2.6
EU-ES	KAL IND	IND	12.07	29.2	+0.9
EU-ES	KAL ALL	IND	11.78	29.4	+1.1
EU-ES	KAL IND	ALL	12.07	34.7	+2.6
EU-ES	KAL ALL	ALL	**11.78**	**34.7**	+2.7
EU-ES	KAL IND	FT	12.07	33.7	+2.5
EU-ES	KAL ALL	FT	11.78	33.9	+2.6
ES-EU	E2E IND	IND	8.26	20.6	+1.3
ES-EU	E2E ALL	IND	8.15	20.6	+1.3
ES-EU	E2E IND	ALL	8.26	22.0	+1.0
ES-EU	E2E ALL	ALL	8.15	22.0	+1.1
ES-EU	E2E IND	FT	8.26	21.5	+1.3
ES-EU	E2E ALL	FT	8.15	21.5	+1.2
ES-EU	KAL IND	IND	7.23	20.9	+1.4
ES-EU	KAL ALL	IND	7.21	20.9	+1.3
ES-EU	KAL IND	ALL	7.23	22.5	+1.2
ES-EU	KAL ALL	ALL	**7.21**	**22.7**	+1.5
ES-EU	KAL IND	FT	7.23	21.9	+1.2
ES-EU	KAL ALL	FT	7.21	22.0	+1.4

Overall, cascaded models trained on all data performed significantly better than their in-domain counterparts, with improvements of up to 5 and 1.6 BLEU points for EU-ES and ES-EU, respectively. These results were mostly due to improvements obtained on the MT components, as was expected from adding significantly larger amounts of training data to the small in-domain datasets. For the ASR components, the impact in terms of WER was minor, with around .3 gains in either language, mainly due to the use of the same data for acoustic modelling in all cases.

Punctuation had a significant impact on the results, with systematic improvements of up to 2.6 and 1.5 BLEU points in EU-ES and ES-EU, respectively. This trend is not entirely surprising, since the translation models were trained on data that include punctuation marks; the impact of punctuation was amplified for models trained on larger amounts of data.

Regarding the overall translation quality, as measured in terms of BLEU scores at least, the results are in line or higher than typical results in similar tasks [11]. One explanation for higher marks is the domain specificity of the corpus, with recurrent topics and typical expressions. Nonetheless, the corpus also features challenging characteristics for automated speech translation, such as the use of Basque dialects or the idiosyncratic properties of the two languages at hand.

From the previous evaluation, we selected the best cascaded variants based on either in-domain or all data and compared with the end-to-end speech translation models, in both translation directions. The comparative results on the mintzai-ST test sets are shown in Table 4, where BP indicates the brevity penalty computed within the BLEU metric.

Table 4: *Results on cascaded and end-to-end models*

LANG	MODEL	ASR	MT	WER	BLEU	BP
EU-ES	CAS	IND	IND	12.07	29.2	0.913
EU-ES	CAS	ALL	ALL	**11.78**	**34.7**	0.978
EU-ES	E2E	-	-	-	17.0	1.000
ES-EU	CAS	IND	IND	7.23	20.9	0.954
ES-EU	CAS	ALL	ALL	**7.21**	**22.7**	0.969
ES-EU	E2E	-	-	-	12.9	1.000

The most notable result from this evaluation is the large difference in terms of BLEU between the cascaded and the end-to-end variants under similar conditions, i.e. using only the in-domain data. Under these conditions, the end-to-end variant was outperformed by 12.2 and 8 BLEU points in EU-ES and ES-EU, respectively. Since the conditions were similar, with relatively small amounts of training data, this large gap may be attributed to the relative dependency of the end-to-end model on larger volumes of training data. Given the noticeably better results obtained with cascaded end-to-end components, alternative end-to-end ST approaches that increase in terms of robustness in low-resource scenarios will need to be further explored.

Interestingly, the end-to-end model produces translation which are closer in length to the human references, as shown by results in terms of brevity penalty. Although further analyses of these aspects will be warranted, these results indicate that the end-to-end systems built for these experiments may be modelling aspects of the speech translation process which are not fully captured by their cascaded counterparts.

5. Conclusions

We described mintzai-ST, the first publicly available corpus for speech translation in Basque-Spanish, shared with the community to support research in the field. The corpus enables development and evaluation of end-to-end or cascaded ST models, and we presented the first results along these lines for this challenging language pair.

In future work, we will prepare variants of the corpus with varying alignment constraints, to measure the impact of larger amounts of training data, in particular for end-to-end models. We will also carry additional analyses of the relative strengths and weaknesses of different ST architectures, building upon the preliminary results described in this paper. Finally, we will explore speech-speech variants of this corpus, by means of speech synthesis, to provide further support to research on speech-to-speech translation.

6. Ackowledgments

This work was supported by the Department of Economic Development and Competitiveness of the Basque Government under project MINTZAI (KK-2019/00065). We wish to thank the anonymous IberSpeech reviewers for their insightful feedback.

7. References

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the International Conference on Learning Representations*, 2015.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010.

[3] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of the IEEE international conference on acoustics, speech and signal processing*, 2013, pp. 6645–6649.

[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 4960–4964.

[5] H. Ney, "Speech translation: Coupling of recognition and translation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999, pp. 517–520.

[6] E. Matusov, S. Kanthak, and H. Ney, "On the integration of speech recognition and statistical machine translation," in *Proceedings of the Ninth European Conference on Speech Communication and Technology*, 2005, pp. 3176–3179.

[7] G. Kumar, G. Blackwood, J. Trmal, D. Povey, and S. Khudanpur, "A coarse-grained model for optimal coupling of ASR and SMT systems for speech translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1902–1907.

[8] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 949–959.

[9] A. Bérard, O. Pietquin, L. Besacier, and C. Servan, "Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation," in *Proceedings of NIPS Workshop on end-to-end learning for speech and audio processing*, 2016.

[10] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," in *Proceedings of INTERSPEECH*, 2017, pp. 2625–2629.

[11] J. Niehues, R. Cattoni, S. Stüker, M. Negri, M. Turchi, E. Salesky, R. Sanabria, L. Barrault, L. Specia, and M. Federico, "The IWSLT 2019 Evaluation Campaign," in *Proceedings of the 16th International Workshop on Spoken Language Translation*, 2019.

[12] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 2012–2017.

[13] J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan, "Europarl-st: A multilingual corpus for speech translation of parliamentary debates," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 8229–8233.

[14] A. Pérez, J. M. Alcaide, and M.-I. Torres, "Euskoparl: a speech and text spanish-basque parallel corpus," in *Proceedings of INTERSPEECH*, 2012, pp. 2362–2365.

[15] T. Etchegoyhen, E. Martínez Garcia, A. Azpeitia, G. Labaka, I. Alegria, I. Cortes Etxabe, A. Jauregi Carrera, I. Ellakuria Santos, M. Martin, and E. Calonge, "Neural Machine Translation of Basque," in *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, 2018, pp. 139–148.

[16] R. Agerri, M. Cuadros, S. Gaines, and G. Rigau, "OpeNER: Open polarity enhanced named entity recognition," *Procesamiento del Lenguaje Natural*, no. 51, pp. 215–218, 2013.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *Proceedings of IEEE Workshop on automatic speech recognition and understanding*, 2011.

[18] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 2007, pp. 177–180.

[19] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy, "Parallel corpora for medium density languages," in *Proceedings of Recent Advances in Natural Language Processing*, 2005, pp. 590–596.

[20] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proceedings of the International Conference on Machine Learning*, 2016, pp. 173–182.

[21] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of INTERSPEECH*, 2019, pp. 2613–2617.

[22] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Proceedings of INTERSPEECH*, 2018, pp. 3743–3747.

[23] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proceedings of INTERSPEECH*, 2015, pp. 3586–3589.

[24] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proceedings of INTERSPEECH*, 2015.

[25] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 2011, pp. 187–197.

[26] O. Tilk and T. Alumäe, "Bidirectional recurrent neural network with attention mechanism for punctuation restoration," in *Proceedings of INTERSPEECH*, 2016.

[27] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in C++," in *Proceedings of ACL 2018, System Demonstrations*, 2018, pp. 116–121.

[28] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1715–1725.

[29] M. A. Di Gangi, M. Negri, and M. Turchi, "Adapting transformer to end-to-end spoken language translation," *Proceedings of INTERSPEECH*, pp. 1133–1137, 2019.

[30] M. A. Di Gangi, M. Negri, R. Cattoni, R. Dessi, and M. Turchi, "Enhancing transformer for end-to-end speech-to-text translation," in *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, Aug. 2019, pp. 21–31.

[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Confidence Measures for Interactive Predictive Neural Machine Translation

Ángel Navarro[1], *Francisco Casacuberta*[1]

[1]Universitat Politècnica de València

annamar8@prhlt.upv.es, fcn@prhlt.upv.es

Abstract

Confidence Measures (CMs) can be used to estimate the reliability of the words of a hypothesis generated by a machine translation system. In the Interactive-Predictive Machine Translation (IPMT) paradigm, they are used to determine which words of the generated predictions need to be corrected, reducing the total number of words typed by the user. The CMs used must be fast enough in order to not affect the interaction between the user and the machine negatively. In this paper, we present several fast CMs for Interactive-Predictive Neural Machine Translation: IBM Model 1 and 2, Fast Align and Hidden Markov Model. These estimators let the system to achieve a reduction in the number of words typed by getting less-quality translations. The experiments done proved that these CMs are fast enough to use them in an IPMT system, and obtained a high relative reduction on the number of words corrected while getting good-quality translations.

Index Terms: confidence measure, neural machine translation, interactive machine translation, interactive predictive machine translation

1. Introduction

Although the quality of the translations generated by Machine Translation (MT) systems has highly improved in recent years with the apparition of the Neural Models, the MT systems are not able to generate error-free translations yet. The Interactive-Predictive Machine Translation (IPMT) field uses human experts to translate interactively with the system the sentences provided, where the machine guarantees high-efficiency and the human the quality of the translations. There are a large variety of approaches that reduce the effort done in the process, one of them is the use of Confidence Measures (CMs).

CMs provide a correctness estimation for each word of a hypothesis. The system uses this information to classify the words as correct if their confidence is above a threshold. The words classified as correct do not need to be checked by the user or corrected, reducing the number of words typed as well the time that the user spends on reading and deciding whether to accept a prediction.

Not all the CMs approaches are adequate for the IPMT paradigm, some of them use large numbers of features, or use the N-best translations for its calculation. In this field, we want to obtain the estimation value as accurate and quick as possible in order to not interrupt the translation process.

2. Related Work

Confidence Estimation has been extensively studied in the Speech Recognition field [1] and opened to MT in the last decade. Blatz et al. (2004) [2] introduced various methods to determine the correctness of the translations based on Statistical Machine Translation (SMT) model and target language features, translation tables and word posterior probabilities. A new method to measure confidence measures is presented by Bach et al. (2011) [3] with a representation of how to visualize the confidence estimations in a post-editing framework.

Recently, these measures have been implemented in IPMT systems [4]. The works presented by González et al. (2010) [5, 6] in a Interactive-Predictive Statistical Machine Translation (IPSMT) system use these confidences based on IBM Model 1 to reduce the number of words to correct by the user, only checking the words with a confidence estimation lower than the threshold set. The workbench CasMaCat [7] shows the CMs to the user using different colours and only displays the predictions up to the first word classified as incorrect.

In this work, we have implemented in an Interactive-Predictive Neural Machine Translation (IPNMT) system four different CMs at word level with the aim of reducing the number of words that the user has to check risking the less quality as possible.

3. Confidence Measures

We have studied different CMs which main features can be pretrained and saved in data matrixes. The system only has to access the matrixes to obtain the confidence estimation, getting the values very fast during the search, which is crucial to do not interrupt the user-machine interaction. The features used are based on the translation probability of the target word and its alignment probability.

The project has focused on the use of computationally efficient CMs over getting high-quality confidence estimations of the words.

3.1. IBM Model 1

The first CM is based on the IBM Model 1 [8], similar to the one described in Blatz et al. 2004 [2]. As performed in related works [5, 6], we modified this CM by replacing the average with the maximal lexicon probability for its dominance over it [9]. Having the sequence $e_1^I = e_1, ..., e_I$ from the target language, and the sequence $f_1^J = f_1, ..., f_J$ from the source langugae, the confidence value of the word e_i can be calculated as follows:

$$c(e_i) = \max_{0 \leq j \leq J} \mathrm{p}(e_i \mid f_j) \tag{1}$$

where $\mathrm{p}(e_i \mid f_j)$ is the lexicon probability obtained from the IBM Model 1, and f_0 is the empty source word.

3.2. IBM Model 2

The second CM is based on the IBM Model 2 [8] and extends the previous CM by adding the alignment probabilities. This extra information lets the system take cognizance of where words appear in either string. The confidence value of the word e_i, which is positioned at i in the target sequence, can be calculated as follows:

10.21437/IberSPEECH.2021-42

$$c(e_i) = \max_{0 \le j \le J} \mathrm{p}(e_i \mid f_j)\ \mathrm{p}(a_i = j \mid i, J, I) \quad (2)$$

where a_i is the alignment position from the source sequence corresponding to position i from target, $\mathrm{p}(a_i|\ i, I, J)$ is the alignment probability obtained from the IBM Model 2, and $a_i = 0$ represents the empty source word.

3.3. Fast Align

The third CM that we have studied is based in Fast Align [10], which appeared as a simpler and faster reparameterization of the IBM Model 2, and presents a different method to calculate the alignment probability of the confidence estimation. In the previous model, we have to compute the alignment probability of each position of the target sentence, alignment and sentence length. In Fast Align this probability is based on favour the alignment points close to the diagonal, and just need to train two parameters, the null alignment probability p_0 and a precision $\lambda \geq 0$ which controls how strongly the model favours the alignment points near the diagonal. The alignment probability can be calculated as follows:

$$\mathrm{p}(a_i{=}j \mid i, J, I) = \begin{cases} p_0 & a_i = 0 \\ (1 - p_0) \times \frac{e^{\lambda h(a_i, i, J, I)}}{Z_\lambda(i, J, I)} & 0 < a_i \leq n \\ 0 & otherwise \end{cases} \quad (3)$$

where $h(a_i, i, J, I)$ can be computed as follows:

$$h(a_i, i, J, I) = -\left| \frac{i}{I} - \frac{a_i}{J} \right| \quad (4)$$

the normalization $Z_\lambda$term is computed as follows:

$$Z_\lambda(i, J, I) = \sum_{j'=1}^{n} \exp \lambda h(j', i, J, I) \quad (5)$$

In their paper [10], Dyer et al. (2013) described in detail how to reduce the time complexity of the method to 1, drastically reducing computing time and obtaining the same time complexity that we had with the previous methods where we only had to get the value from a matrix.

3.4. Hidden Markov Model

The last CM is based in HMM [11], which differs from the previous CMs by taking a different approach to obtain the alignment probabilities. HMM does not take in count the position of the target word, its alignment probability is calculated from the alignment positions on the source sentence of a target word and the previous one, more specifically it depends only on the jump width $(a_i - a_{i-1})$. The confidence value can be calculated as follows:

$$c(e_i) = \max_{0 \le j \le J} \mathrm{p}(e_i \mid f_j)\ \mathrm{p}(a_i{=}j|\ a_{i-1}, J) \quad (6)$$

where $\mathrm{p}(a_i{=}j|\ a_{i-1}, J)$ is the alignment probability obtained from HMM that can also be represented as $p(j|j', J)$. To obtain the confidence value of a target word the method requires to calculate the optimal alignment of the previous word. This requires to use dynamic programming as follows:

$$\hat{a}_{i-1} = \arg\max_{1 \le a_{i-1} \le J} p(a_i|a_{i-1}, J) Q(i-1, a_{i-1}) \quad (7)$$

where $Q(i, j)$ is a sort of partial probability that we can calculate recursively using the following formula:

$$Q(i, j) = \mathrm{p}(e_i|f_j) \max_{1 \le j' \le J} \left[p(j|j', J)\ Q(i-1, j') \right] \quad (8)$$

To calculate it efficiently, we need to save in memory all the partial probabilities of the previous target positions. This information is used recursively for each partial probability, so we are saving computation time by keeping it on memory.

4. Experimental Setup

4.1. System Evaluation

As explained previously, CMs are a classifying task where we want to tag the words of the hypothesis generated by the system as correct or incorrect, depending on its confidence value and the threshold used to decide correctness. The metrics used to evaluate the CMs, Classification Error Rate (CER) and Receiver Operating Characteristic (ROC), capture the discriminability of the classification function across the range of all thresholds used. We are also going to report the mean execution time of the confidence measures in milliseconds.

The ground truth of the words classified by the models is obtained from the reference translations of the parallel corpus. We consider a word as correct if it occurs in the same position as the reference translation. As we are highly restricting the number of words that could be classified as correct the metrics obtained are pessimistic.

The CER [5] is computed as the proportion of words that our CM has classified incorrectly given a threshold value. The area under the ROC curves [12], called IROC, gives us a global indication of the CM discriminability. These curves show the plot correct-reject ratio (true correct / n_0) vs correct-accept ratio (true incorrect / n_1) for different thresholds, where n_0 and n_1 are the total number of correct and incorrect words of the ground truth. The ROC curve lies in the unit square, the diagonal corresponding to the random choice and the edges to a perfect classification.

To evaluate the improvement of the IPNMT system with the CM implementation, we compare the improvement on BiLingual Evaluation Understudy (BLEU) with the reduction of Word Stroke Ratio (WSR) [13]. BLEU computes a geometric mean of the precision of n-grams multiplied by a factor to penalise short sentences. WSR is computed as the proportion of words that the user needs to correct to generate the reference translation.

4.2. Corpora

All experiments have been carried out on the Spanish-English language pair of the EU corpus. The corpus was cleaned, lowercased and tokenized using the scripts included in the toolkit Moses [14]. We applied the subword subdivision BPE, described in Sennrich et al. [15], with a maximum of 32000 merges.

The EU corpus [16] is formed from the Bulletin of the European Union, which exists in all official languages and is publicly available on the internet.

4.3. Experimental Setup

First of all, we built our Neural Machine Translation (NMT) models using NMT-Keras [17]. We used an encoder-decoder

Table 1: *Statistics of the Spanish-English EU corpus. K and M stands for thousands and millions respectively.*

		Es-En	
Training	Sentences	214K	
	Average Length	27	24
	Running Words	6M	5M
	Vocabulary	84K	69K
Dev.	Sentences	400	
	Average Length	29	25
	Running Words	12K	10K
Test	Sentences	800	
	Average Length	28	25
	Running Words	23K	20K

architecture with attention model [18] and LSTM cells [19]. The dimensions of encoder, decoder, attention model and word embedding were set to 512. We used a single hidden layer of encoder and decoder. The learning algorithm used for the NMT system was Adam [20], with a learning rate of 0.0002. We clipped the L_2 norm of the gradient to 5. The batch size was set to 50 and the beam size to 6.

Secondly, we built our Confidence Measures Models. We use the toolkit GIZA++ [21] to train the IBM Model 1 and 2; and the HMM Model. To built the Fast Align Model we used the scripts developed in Dyer et al. (2013) [10].

4.4. Confidence Measures Evaluation Results

We carried out experimentation intended to study the performance of the CM on an IPNMT system. First of all, we carried out an IPMT session that we used to produce a corpus of words tagged as correct or incorrect. These words are compared with the references to classify them correctly and use them as the ground truth to calculate the CER and ROC.

Figure 1 displays the CER evolution through different threshold values for each one of the CMs used. The three models that used an alignment probability for their confidence calculation have similar behaviour. The IBM Model 2 obtained the best CER score, 0.24 for a threshold value of 0.125.

Figure 2 compares the ROC curves of the CMs used, the diagonal shows the random choice curve. This time the behaviour of the IBM Model 2 and Fast Align are very similar related, though Fast Align is some points lower. At the same time, the IBM Model 1 behaves like HMM.

Table 2 shows the performance of the confidence measures in terms of CER and IROC. The baseline is a classifier which tags all the words as the most frequent class, $\text{CER}_b = min(n_0, n_1)/n$. The values of CER displayed are based on threshold optimized on the validation set. All the CMs obtain a relative improvement over the baseline CER of more than 7%. The best CM is the model based on the IBM Model 2 that gets a relative improvement over the baseline CER of 20%.

All the CMs obtained an execution time lower than 100 milliseconds, a threshold set by Nielsen (1994) [22] that marks the limit for having the user feel that the system is reacting instantaneously. This makes the confidence measures optimal for an IPMT system and do not break the human-machine interaction.

4.5. User Simulated IPNMT Results

In the previous section, we have studied the discriminability of the different confidence measures across the range of all thresholds used. We have integrated these confidence measures in an IPNMT system to study the trade-off between the effort that the translator needs to do and the final quality of the translations.

For the user simulation, we only check and correct those words that have a confidence estimation under the threshold value. The words are compared with the ones that have the same position on the reference sentence and are corrected if they are different. We correct the words typing the ones that appear on the reference without taking in count the context.

In this section, we present a range of experiments using different thresholds values from 0.0 where the system behaves as an unsupervised NMT system, to 1.0 where the user has to check and correct all the words as an IPNMT system. For each threshold used, we compare the user effort using the metric WSR, and the quality of the sentences with BLEU.

Figure 3 shows the WSR and BLEU scores for all the CM used across the transition between an unsupervised NMT system with a 0.0 threshold and the conventional IPNMT system with a 1.0 threshold. As we raise the threshold more words are tagged as incorrect increasing the number words that the user has to check and correct, which improves the quality of the translations. Although IBM Model 1 and HMM obtained the lowers IROC values, they present more gradual transitions that let us have a larger range of useful thresholds values to use.

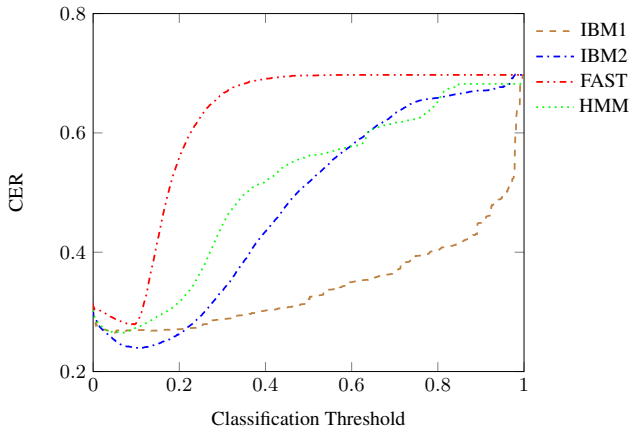


Figure 1: *CER for IBM Model 1, IBM Model 2, Fast Align and HMM across the range of all thresholds used.*

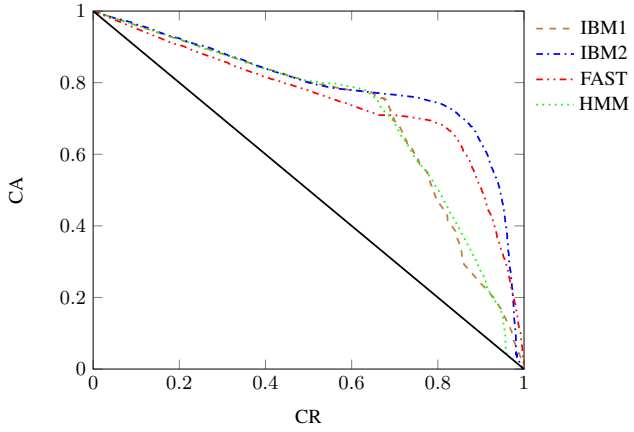


Figure 2: *ROC curves for IBM Model 1, IBM Model 2, Fast Align and HMM.*

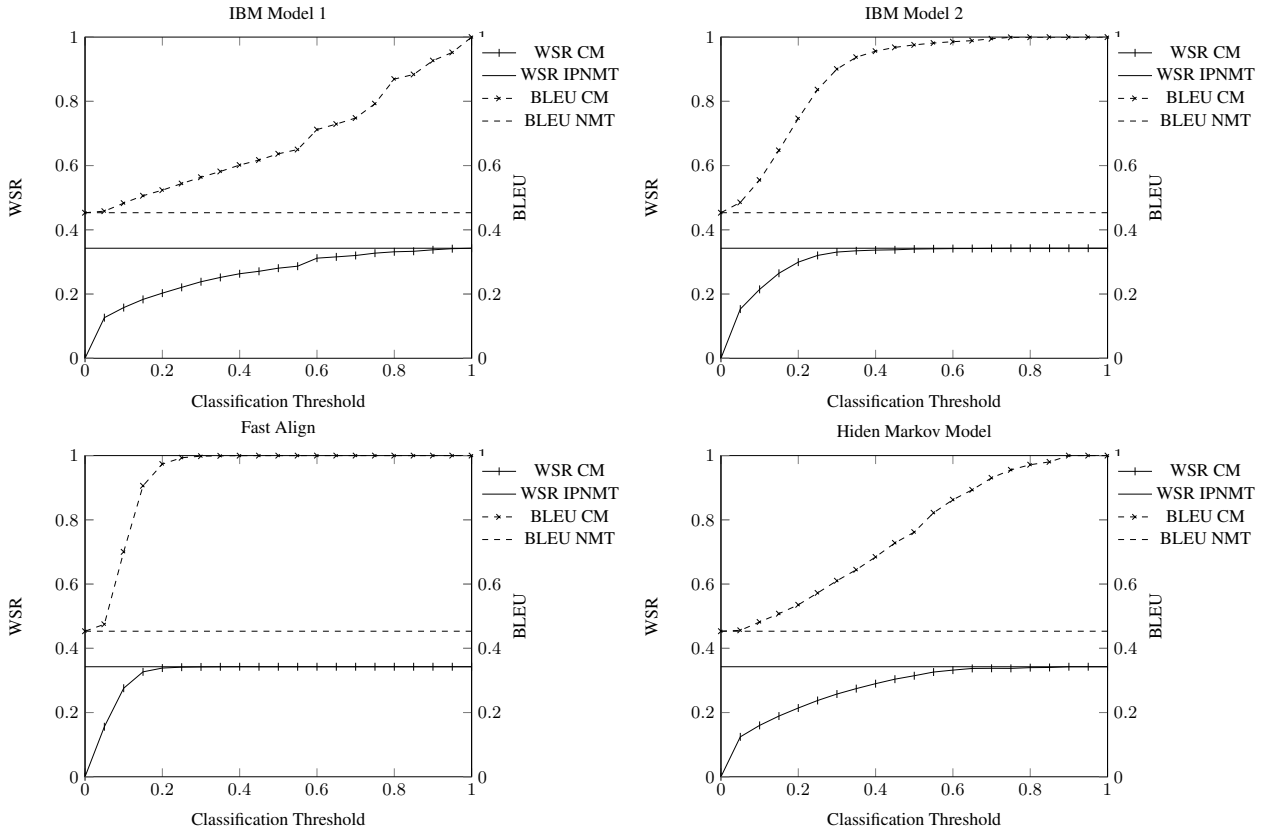


Figure 3: *BLEU translation scores versus WSR across the range of all thresholds used.*

Table 2: *CER[%], IROC and execution time (ms) of the Confidences Measures on the test set.*

Confidence Measure	CER	IROC	ms
baseline	30.0	-	-
IBM Model 1	26.5	0.713	**3.41**
IBM Model 2	**23.9**	**0.791**	4.24
Fast Align	27.8	0.752	4.32
HMM	26.5	0.714	8.93

We can compare the relative WSR reduction obtained for each CM while getting similar quality translations. For this purpose, we compare the relative WSR reductions of the experiment samples with BLEU scores closer to 0.70. The higher relative reduction is obtained by Fast Align with a relative reduction of 19.5%. IBM Model 1, IBM Model 2 and HMM obtained 6.6%, 12.6% and 11.3% respectively.

5. Conclusions and Future Work

5.1. Conclusions

In this paper, we have proposed four different CMs that can be computed very fast while obtaining a good discriminability evaluation, which makes them a perfect option to implement into IPMT systems. We compared the CM using the CER and IROC metrics. The CM based on IBM Model 2 obtained the best results in both metrics.

We have tested the confidence measures in an IPNMT system, comparing the effort that the user has to do for each threshold value used with the quality of the translations obtained. Around 0.70 of BLEU score Fast Align obtained the best WSR reduction, almost 20%.

5.2. Future Work

The word confidence measures obtained can be combined to compute a sentence correctness value. As future work, we plan to investigate different methods to combine them and compare the effort reduction.

Also, we will try in future work to use more complex CMs with higher computational time, like neural models, and try to use them in IPMT systems.

In the experiments that we have performed, we simulated the user interaction and used for the evaluation of very pessimistic ground truth. We need to compare our results with those obtained with real translators that will take into account different possible translations in the correction process.

6. Acknowledgements

This work received funds from the Comunitat Valenciana under project EU-FEDER (*IDIFEDER*/2018/025), Generalitat Valenciana under project ALMAMATER (*PrometeoII*/2014/030), and Ministerio de Ciencia under project MIRANDA-DocTIUM (RTI2018-095645-B-C22).

7. References

[1] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on speech and audio processing*, vol. 9, no. 3, pp. 288–298, 2001.

[2] J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing, "Confidence estimation for machine translation," in *Coling 2004: Proceedings of the 20th international conference on computational linguistics*, 2004, pp. 315–321.

[3] N. Bach, F. Huang, and Y. Al-Onaizan, "Goodness: A method for measuring machine translation confidence," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 211–219.

[4] M. Domingo, A. Peris, and F. Casacuberta, "Segment-based interactive-predictive machine translation," *Machine Translation*, vol. 31, no. 4, pp. 163–185, 2017.

[5] J. González-Rubio, D. Ortíz-Martínez, and F. Casacuberta, "On the use of confidence measures within an interactive-predictive machine translation system," in *Proceedings of the 14th Annual conference of the European Association for Machine Translation*. Saint Raphaël, France: European Association for Machine Translation, May 27–28 2010. [Online]. Available: https://www.aclweb.org/anthology/2010.eamt-1.18

[6] J. González-Rubio, D. Ortiz-Martínez, and F. Casacuberta, "Balancing user effort and translation error in interactive machine translation via confidence measures," in *Proceedings of the ACL 2010 Conference Short Papers*, 2010, pp. 173–177.

[7] V. Alabau, R. Bonk, C. Buck, M. Carl, F. Casacuberta, M. García-Martínez, J. González, P. Koehn, L. Leiva, B. Mesa-Lao *et al.*, "Casmacat: An open source workbench for advanced computer aided translation," *The Prague Bulletin of Mathematical Linguistics*, vol. 100, no. 1, pp. 101–112, 2013.

[8] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993. [Online]. Available: https://www.aclweb.org/anthology/J93-2003

[9] N. Ueffing and H. Ney, "Application of word-level confidence measures in interactive statistical machine translation," in *Proceedings of the 10th EAMT Conference: Practical applications of machine translation*. Budapest, Hungary: European Association for Machine Translation, May 30–31 2005. [Online]. Available: https://www.aclweb.org/anthology/2005.eamt-1.35

[10] C. Dyer, V. Chahuneau, and N. A. Smith, "A simple, fast, and effective reparameterization of IBM model 2," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 644–648. [Online]. Available: https://www.aclweb.org/anthology/N13-1073

[11] S. Vogel, H. Ney, and C. Tillmann, "HMM-based word alignment in statistical translation," in *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*, 1996. [Online]. Available: https://www.aclweb.org/anthology/C96-2141

[12] R. O. Duda, P. E. Hart *et al.*, *Pattern classification*. John Wiley & Sons, 2006.

[13] J. Tomás and F. Casacuberta, "Statistical phrase-based models for interactive computer-assisted translation," in *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 2006, pp. 835–841.

[14] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 177–180. [Online]. Available: https://www.aclweb.org/anthology/P07-2045

[15] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: https://www.aclweb.org/anthology/P16-1162

[16] S. Barrachina, O. Bender, F. Casacuberta, J. Civera, E. Cubel, S. Khadivi, A. Lagarda, H. Ney, J. Tomás, E. Vidal, and J.-M. Vilar, "Statistical approaches to computer-assisted translation," *Computational Linguistics*, vol. 35, no. 1, pp. 3–28, 2009. [Online]. Available: https://www.aclweb.org/anthology/J09-1002

[17] Álvaro Peris and F. Casacuberta, "NMT-Keras: a Very Flexible Toolkit with a Focus on Interactive NMT and Online Learning," *The Prague Bulletin of Mathematical Linguistics*, vol. 111, pp. 113–124, 2018. [Online]. Available: https://ufal.mff.cuni.cz/pbml/111/art-peris-casacuberta.pdf

[18] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2017.

[21] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[22] J. Nielsen, *Usability engineering*. Morgan Kaufmann, 1994.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Sentence Embeddings and Sentence Similarity for Portuguese FAQs

Nuno Carriço[1], *Paulo Quaresma*[2]

[1,2]Computer Science Department, University of Évora

nfmc@uevora.pt, pq@uevora.pt

Abstract

Virtual Assistant Bots are becoming essential in business models. This aims to provide customer service without the need of a human operator. Thus, the first step is to understand what a customer needs. To achieve this, we compute the sentence distance between a set of predefined FAQs and the user sentence, and extract the closest FAQ. While the problem has satisfactory results for English language, it is not the case for Portuguese language. Therefore, we propose the use of portuguese BERT models to obtain the sentence embeddings of both the FAQs and user sentence, in order to compute their distances scores. The BERT models are fine tuned with the ASSIN 2 dataset for sentence similarity tasks to achieve better performance. The fine tuned models were evaluated against ASSIN 2 test set.

The FAQs embeddings are inserted in a FAISS index, which is used to extract the n closest FAQs embeddings to a user sentence. The index provides an efficient way to maintain the embeddings and search for the closest neighbors given a query data point. Given the set of FAQs, we built sample user questions, labelled with their corresponding FAQ, to test the setup.

Index Terms: BERT, sentence embeddings, sentence similarity, paraphrase searching

1. Introduction

Virtual Assistant Bots are becoming an essential part of a company customer service. A customer is able to receive support at any time without the need of directly contact a company's assistance line and wait for a response. Furthermore, these assistants are now evolving in the direction of more personalized support and can be easily deployed to a variety of different tasks. Nonetheless, the first step to all of this is understanding. Our virtual assistant must be capable of understanding, to some degree, what the customer's intention is. For that matter, the collection of most common problems/questions (FAQs) a customer might encounter is a valuable resource to start with. These FAQs will serve as a base to compare user sentences as a way to "understand" what the customer needs. Therefore, we propose a simple and scalable system capable of efficiently computing the closest FAQ to a user sentence.

1.1. Related Work

Word embeddings has been a wide research topic. Various methods of obtaining word embeddings have been successful, such as ELMo [1] , which enriches word embeddings using internal network layer information, or BERT [2] with successive application of attention mechanisms to extract relations between words in a given sentence. Other methods include positional dependency to obtain the embeddings [3], that is, a triplet $(w_t,\ c,\ w_c)$ is used to construct the embeddings, where w_t, w_c, c represent, respectively, the target word, the context word and the positional dependency computed from the context.

Sentence encoders have also been in study and are applied succefuly in numerous natural language processing tasks. These encoders range from simple LSTM networks to self attention networks and hierarchical convolutional networks [4]. Recently, BERT embeddings have been in use to produce meaningful representations of sentences with the assistance of siamese networks [5].

Semantic search engines were taken a step further using word embeddings. [6] used ELMo as a base to build a word embedding based search engine. The procedure is simple and works as follows:

1. Input a query sentence and obtain it's embedding using ELMo.
2. Compute the cosine similarities scores between the previous embedding and the remaining embeddings.
3. Return the n closest vectors.

However, as the number of vectors in the search space grows, the complexity of the search also increases. Similar approaches were taken by [7] which makes BERT available as service. In this case, the embeddings are obtained with BERT, and the scores computed using the normalized dot product between the two embedding vectors. This approach also suffers from the search complexity problem. With our proposed framework, we make use of sentence embeddings to grasp the general meaning of the whole sentence. Also, the navigation trough the search space is done using an index as a way to reduce search speed.

1.2. Framework

The proposed framework consists of two sub tasks. First, we need to represent both the FAQs and user sentences by a vector which encodes their semantic values. Second, given the sentence's representation, we need to compute their similarity. This will allow us to pick the most likely FAQ that satisfies the customer intention.

Regarding the encoding step, a variety of options to build word embeddings were already available, such as Word2vec [8], GloVe [9], ELMo [1] and BERT [2]. However, we intended to encode the whole sentence, not just the relations between the words. For this reason, we decided to use Sentence BERT (SBERT) [5] as our transformer. SBERT computes the sentence embedding in the following manner: first, it computes the word embeddings using a BERT model; to the previous output, it applies a pooling layer to build the sentence embedding. For the pooling layer, we have the following strategies:

- Computing the mean of the output vectors.
- Computing the max of the output vectors.
- Using the CLS token.

Moreover, the underlying BERT model is fine tuned with siamese and triplet networks with the purpose of producing meaningful sentence embeddings that can be easily compared.

10.21437/IberSPEECH.2021-43

Since our proposed framework is expected to attent online requests from customers, we need an efficient way to search the FAQ search space for the closest FAQ to a user sentence using a given metric for similarity. For this step we use FAISS [10]. FAISS is an open source library developed by Facebook AI research group. It is mainly used for efficient space search and clustering of vectors. FAISS builds a data structure, an index, representing a given set of vectors. When a query vector is introduced in the search space, FAISS efficiently computes the distances between the query vector and the remaining index vectors, returning the k-closest index vectors. FAISS makes available multiple types of indexes. For the problem at hand, we chose the indexes IndexFlatL2 and IndexFlatIP, which use the euclidean distance and inner product, respectively, as distance metrics.

A preview of the system's architectures is displayed in figure 1.

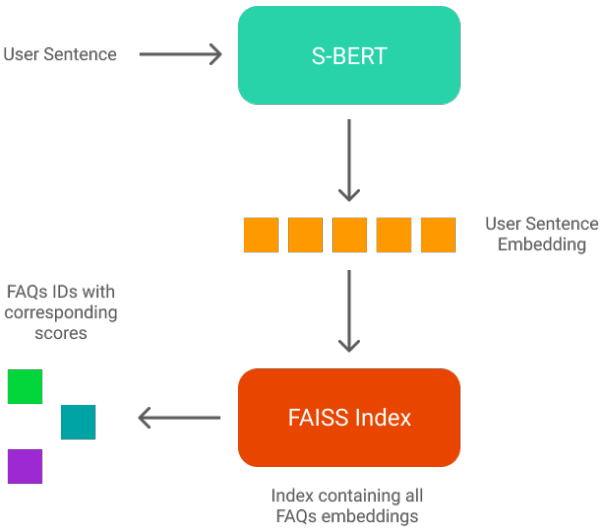


Figure 1: *Proposed Framework Architecture*

1.3. FAQ Corpora

The FAQ corpora was provided by a Portuguese telecommunications company and contains a total of 72 FAQs, ranging over topics related to employee assistance. These topics include matters such as, holidays, supplementary work, family, communication benefits, meals and retirement. Examples of such questions are presented in Table 1.

1.4. ASSIN 2 Dataset

ASSIN 2 [1] corpus consists of simple sentences that do not include named entities or indirect speech and all verbs are in the present tense. The training set is composed of 6500 sentence pairs, while the validation set contains 500 sentence pairs. These sentences are manually annotated for text entailment and sentence similarity. The semantic similarity values range over 1 to 5 and the text entailment classes are one of entailment or none. For the test set, it contains roughly 3000 sentence pairs, following the same annotation schema.

For the evaluation step, F1 of precision and recall is used to evaluate text entailment, and Pearson correlation is applied to the semantic similarity task. Below we provide the scores of the highest performing contestants for the semantic similarity task.

[1] https://sites.google.com/view/assin2/english

Table 1: *Example Questions in FAQ corpora*

Topic	Example Question
Communication Benefits	What are the communication benefits after retiring?
Family	Regarding family assistance, who supports the payment?
Meals	What is the processing rule of meal allowance?
Holidays	Is it allowed to anticipate holidays?
Retirement	How to request pension support?

Table 2: *ASSIN 2 Similary Task Results*

Team	Pearson	MSE
IPR	0.826	0.52
IPR	0.809	0.62
L2F/INESC	0.778	0.52
Stilingue	0.800	0.39
Stilingue	0.817	0.47

2. Experiments

2.1. FAQ Test Set

For the given FAQ corpora, we were not provided with user queries, therefore, we developed a simple set of possible user questions. For each topic in the corpora, we chose two questions and rephrased them, ending with a total of 24 questions. Examples are provided in Table 3. For the purposes of testing, each rephrased question is labelled with its correct/original FAQ.

2.2. Pre Trained

Based on the performance on the FAQ test set, we filtered out pre trained models that wouldn't be worth fine tuning. The pre trained BERT models tested were the following: Portuguese language BERT models [11], namely bert-base-portuguese-cased (BBPC) and bert-large-portuguese-cased (BLPC); English language BERT models, that is, roberta-base-nli-stsb-mean-tokens (Roberta-BNS-Mean), roberta-large-nli-stsb-mean-tokens (Roberta-LNS-Mean) and distilbert-base-nli-stsb-mean-tokens (Distilbert-BNS-Mean).

The procedure worked as follows:

1. Using SBERT, we computed the sentence embeddings of both the user sentences and the FAQs, using HugingFace Transformers [12].

Table 3: *Example Question Rephrasing*

Original FAQ	Rephrased FAQ
What are the conditions to acquire equipment in a store?	How can equipment be acquired?
Is it allowed to anticipate holidays?	Can I anticipate holidays?
How many hours of effective work are necessary in order to request meal allowance?	How many hours do I have to work to request meal allowance?

2. By applying cosine similarity to each pair of user sentence embedding and FAQ embedding, we obtained the similarity scores.
3. Finally, for each user query, we picked the FAQ with the highest similarity score as the predicted FAQ.

To evaluate the results, we compute the accuracy of the predicted FAQs. The preliminary results for the best pre trained models are as follows in Table 4

Table 4: *Preliminary Testing*

Model	Accuracy
BBPC-Mean	**0.875**
BBPC-Max	0.791
BLPC-Mean	0.75
BLPC-Max	0.833
Distilbert-BNS-Mean	0.833
Roberta-BNS-Mean	**0.875**
Roberta-LNS-Mean	0.791

From the results presented in Table 4, the models BBPC-Mean and Roberta-BNS-Mean produce the best results with an accuracy of 0.875. However, since Roberta-BNS-Mean is an English language model, we do not proceed to fine tuned it with the ASSIN 2 dataset.

2.3. Trained

Given that our goal was to build a setup for Portuguese language FAQs, we picked the best Portuguese language models from the preliminary results, namely BBPC-Mean/Max and BLPC-Mean/Max. These models were further fine tuned using ASSIN 2 dataset for 10 epochs using a batch size of 16 with cosine similarity loss as the loss function. We also experimented with different post processing layers, including CNN, LSTM and DAN.

Testing these fine tuned models against the FAQ test set, we got the results presented in Table 5.

Based on the results, we chose one model for each of the BBPC and BLPC base models, in this case, BBPC-Mean-CNN and BLPC-Mean. Models without post processing were favoured over those with post processing layers, in case of equal accuracy. These two models were further tested against ASSIN 2 test set, yielding the results in Table 6.

Table 5: *Trained Models Results*

Model	No Post Processing	CNN	LSTM	CNN-DAN
BBPC-Mean	0.875	**0.917**	0.833	0.875
BBPC-Max	**0.875**	0.875	0.833	0.875
BLPC-Mean	**0.917**	0.917	0.875	0.875
BLPC-Max	0.75	**0.875**	0.833	0.875

Table 6: *ASSIN 2 Comparison*

Model/Team	Pearson	MSE
BBPC-Mean-CNN	0.822	0.8
BLPC-Mean	**0.831**	0.78
IPR	0.826	0.52
IPR	0.809	0.62
L2F/INESC	0.778	0.52
Stilingue	0.800	**0.39**
Stilingue	0.817	0.47

Looking at Table 6, we observe that the fine tuned model BLPC-Mean outperformed the best contestant by a factor of 0.05. Nonetheless, the model BBPC-Mean-CNN also keeps on pair with the majority of the remaining contestants. Both of these could be improved to reduce the MSE scores, which are the highest among the given results.

2.4. FAISS Indexes

The setup is required to run online, that involves computing efficiently the most similar FAQ to a user query. Moreover, we kept in mind that FAQs change over time, so it could necessary to add or remove FAQs from the search space. For that matter, we introduce FAISS indexes to our framework.

First, we compute all the FAQs sentence embeddings and store them in a FAISS index. In case of needing a new FAQ, we simply coompute its sentence embedding and insert it into the index. The removal is just as easy, since we can attribute an id to each vector in the index, we can simply remove an embedding using its corresponding id.

When a new user sentence arrives, it is computed its sentence embedding and we query the index for the k closest FAQs embeddings, returning it's ids.

To test this approach, we used the developed FAQ set together with two FAISS indexes: IndexFlatL2 and IndexFlatIP. The results are shown in Table 8

Table 7: *Sample Results*

Query	BBPC-Mean-CNN	BLPC-Mean	BBPC-Mean-CNN IndexFlatIP	BLPC-Mean IndexFlatL2
Posso usufruir da dispensa semanal em cursos sem componente letiva?	Qual o código para gozo de férias por trabalho suplementar? ✗	Qual o código para gozo de férias por trabalho suplementar? ✗	Qual o código para gozo de férias por trabalho suplementar? ✗	Qual o código para gozo de férias por trabalho suplementar? ✗
Como pedir benefícios de comunicações?	Como requerer/transferir benefícios comunicações colaborador para uma nova conta/linha de rede (telefone)? ✓	Quais os benefícios de comunicações na reforma? ✗	Os trabalhadores não ativos podem usufruir do cartão Galp frota colaboradores? ✗	Como requerer/transferir benefícios comunicações colaborador para uma nova conta/linha de rede (telefone)? ✓
Que taxa de IRS aplica a empresa?	Qual o valor das ajudas de custo nacional/estrangeiro? ✗	Que taxa de IRS me foi aplicada pela empresa? ✓	Que taxa de IRS me foi aplicada pela empresa? ✓	Que taxa de IRS me foi aplicada pela empresa? ✓

Table 8: *Index Results*

Model	IndexFlatL2	IndexFlatIP
BBPC-Mean	0.875	0.833
BBPC-Mean-CNN	0.875	**0.917**
BBPC-Max	0.833	0.5
BBPC-Max-CNN	0.875	0.208
BLPC-Mean	**0.958**	**0.917**
BLPC-Mean-CNN	0.917	0.875
BLPC-Max	0.75	0.875
BLPC-Max-CNN	0.833	0.667

From Table 8, for the IndexFlatL2 index, the model BLPC-Mean outperformed the remaining models with an accuracy of 0.958. Focusing our attention on the IndexFlatIP index, the best performing models are BBPC-Mean-CNN and BLPC-Mean, both having an accuracy of 0.917. Based on these results, the model BLPC-Mean together with IndexFlatL2 produces the best setup.

2.5. Results

For the developed test set, the models, in general, select the correct FAQ in the majority of the cases, keeping the accuracy above 80%, as we already presented in tables 5 and 8. However, there are some sentences for which the models did not find the similarity. Looking at Table 7, we depicted the most problematic queries. For the first query, no model found the correct FAQ, which, in this case, was the text *Nos cursos em que não existe componente letiva (frequência de aulas), designadamente, em mestrados e doutoramentos é possível beneficiar da dispensa semanal prevista no estatudo de trabalhador - estudante ?*. The lack of fine tuning with FAQ domain data could be causing this issue, resulting in distant embeddings for this query and respective FAQ. Regarding the two other queries, the use of indexes generally helps finding the correct FAQ, specially, the model BLPC-Mean for the second query, and the model BBPC-Mean-CNN for the third query.

It would be needed a much larger dataset to assess how the system scales. Nonetheless, as of the time of writing, the system is currently being tested in such scenarios by the requesting company, and soon we will be able to provide results on that matter.

3. Conclusion

With this paper, we aimed at developing a framework that could serve as a baseline for online sentence similarity search, in this case, for virtual assistant bots and FAQs similarity. For the Portuguese language, the results show that the BLPC-Mean model together with the IndexFlatL2 index are the more promising combination for the task at hand. Nonetheless, this framework can be extended to any language simply by switching the underlying SBERT model to a model of the desired language.

Future work could involve applying dimensional reduction techniques and testing new underlying models using recent research, such as tBERT. Additionally, considering the FAQ's answers to find the correct FAQ would be of interest as well. Moreover, instead of using FAISS indexes, it could be of use trying different clustering methods to extract the closest FAQ to a sentence. Also, it would be useful to develop a full Portuguese language dataset for FAQ search as a base reference for future projects.

4. References

[1] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-

training of deep bidirectional transformers for language understanding," 2019.

[3] Y. Yin, C. Wang, and M. Zhang, "Pod: Positional dependency-based word embedding for aspect term extraction," 2020.

[4] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," 2018.

[5] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[6] "Elmo: Contextual language embedding," https://towardsdatascience.com/elmo-contextual-language-embedding-335de2268604, accessed: 2020-12-20.

[7] "Bert as a service," https://github.com/hanxiao/bert-as-service#building-a-qa-semantic-search-engine-in-3-minutes, accessed: 2020-12-20.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[9] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[10] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *arXiv preprint arXiv:1702.08734*, 2017.

[11] F. Souza, R. Nogueira, and R. Lotufo, "BERTimbau: pretrained BERT models for Brazilian Portuguese," in *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020.

[12] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Domain Adaptation in Dialogue Systems using Transfer and Meta-Learning

Rui Ribeiro[12], *Alberto Abad*[12], *José Lopes*[3]

[1]INESC-ID Lisboa, Portugal
[2]Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3]Heriot-Watt University, Edinburgh, United Kingdom

rui.m.ribeiro@tecnico.ulisboa.pt

Abstract

Current generative-based dialogue systems are data-hungry and fail to adapt to new unseen domains when only a small amount of target data is available. Additionally, in real-world applications, most domains are underrepresented, so there is a need to create a system capable of generalizing to these domains using minimal data. In this paper, we propose a method that adapts to unseen domains by combining both transfer and meta-learning (DATML). DATML improves the previous state-of-the-art dialogue model, DiKTNet, by introducing a different learning technique: meta-learning. We use Reptile, a first-order optimization-based meta-learning algorithm as our improved training method. We evaluated our model on the MultiWOZ dataset and outperformed DiKTNet in both BLEU and Entity F1 scores when the same amount of data is available.
Index Terms: dialogue systems, domain adaptation, transfer-learning, meta-learning

1. Introduction

With the appearance of chatbots like Siri and Alexa capable of having fluent and consistent conversations, dialogue systems have become very popular these days. Additionally, the emergence of deep learning techniques in natural language processing contributes to this popularity and various new models were created in order to surpass previous rule-based models. However, these generative-based models are data-hungry, they need large amounts of training data in order to obtain good results, they produce dull responses and fail to adapt to new unseen domains when only a few examples of data are available. Besides, in real-world applications, most domains are underrepresented, so there is a need to create a model capable of generalizing to these domains using the minimum amount of data available.

In this paper, we study the importance of generalizing to unseen domains using minimal data and aim to design a novel model to surpass this problem. We believe that for successful adaptation to new domains, two key features are essential for improving the overall performance of a dialogue system: better representation learning and better learning techniques. Following this belief, we are concerned with the exploration of a method able to learn a more general dialogue representation from a large data-source and able to incorporate this information into a dialogue system.

We follow this reasoning and introduce Domain Adaptation using Transfer and Meta-Learning (DATML), a model that combines both transfer-learning with meta-learning for the purpose of adapting to unseen domains. Our model builds upon the approach from Dialogue Knowledge Transfer Network (DiKTNet) [1] by enhancing its learning method while keeping the strong representation learning present in both ELMo [2] contextual embeddings and latent representations. For that, we divide the training method into three training stages: 1. A pre-training phase where some latent representations are leveraged from a domain-agnostic dataset; 2. Source training with all data except dialogues from the target domain; 3. Fine-tuning using few examples from the target domain.

We incorporate meta-learning in source training as this method proved to be promising at capturing domain-agnostic dialogue representations [3]. However, instead of using Model-Agnostic Metal-Learning (MAML) [4] algorithm, we use a first-order optimization-based method, Reptile [5], which has shown to achieve similar or even better results than MAML for low-resource NLU tasks while being more lightweight in terms of memory consumption [6].

We evaluate our model on the MultiWOZ dataset [7] and compare our approach with both Zero-Shot Dialog Generation (ZSDG) [8] and current state-of-the-art model in few-shot dialogue generation, DiKTNet. As the code for both baselines is openly available online, we adapt and evaluate their implementations on the MultiWOZ corpus. Our model outperforms both ZSDG and DiKTNet when the same amount of data is available. Furthermore, DATML achieves superior performance with 3% of available target data in comparison to DiKTNet with 10%, which shows that DATML surpasses DiKTNet in terms of both performance and data efficiency.

2. Related Work

The reduced amount of available data has always been a problem in domain adaptation tasks. Methods as meta-learning [4], transfer-learning [9, 10, 11] and few-shot learning [12, 13, 14] were introduced to solve this problem in machine learning. However, there were only a few attempts to solve the problem of domain adaptation in end-to-end dialogue systems.

Perhaps, one of the first studies to pursue this direction was the work from ZSDG [8], where authors performed zero-shot dialogue generation using minimal data in the form of seed responses. The model is described as "zero-shot" and does not use complete dialogues, however, the model still depends on human annotated data. Although this approach seems promising, ZSDG relies on these annotations for seed responses, and in the real-world scenario, if collecting data for underrepresented domains is already difficult enough, access to annotated data becomes infeasible.

More recent studies attempt to perform domain adaption without the need of human annotated data and adopt the methods presented above: Domain Adaptive Dialog Generation via Meta-Learning (DAML) [3] incorporates meta-learning into the *sequicity* [15] model to train a dialogue system able to generalize to unseen domains. This approach seems promising, yet DAML was evaluated on a synthetic dataset. DiKTNet [1] applies transfer learning by leveraging general latent representa-

10.21437/IberSPEECH.2021-44

tions from a large data-source and incorporating them into a Hierarchical Recurrent Encoder-Decoder (HRED). We will describe this model in detail in the following sections as it represents a key feature for our solution.

3. Base Model

As mentioned in the previous section, our base model is the work from DiKTNet [1]. The basic idea in DiKTNet is learning reusable latent representations from a domain-agnostic dataset and incorporate that knowledge when training using minimal data from the target domains. DiKTNet's base model is the same from ZSDG, a HRED with an attention-based copying mechanism.

More formally, the base model's HRED $\mathcal{F}$ is optimized according to the following loss function:

$$\mathcal{L}_{HRED} = \log p_{\mathcal{F}^d}(\mathbf{x}_{sys}|\mathcal{F}^e(\mathbf{c}, \mathbf{x}_{usr})), \tag{1}$$

where $\mathbf{x}_{usr}$ is the user's request, $\mathbf{x}_{sys}$ is the system's response and $\mathbf{c}$ is the context.

Although each domain has its specific dialogue structure, every domain still shares a general representation. Thus, the authors consider the Latent Action Encoder-Decoder (LAED) framework [16]. LAED is, in essence, a Variational Auto-Encoder (VAE) representation method that allows discovering interpretable and meaningful representations of utterances into discrete latent variables. LAED introduces a recognition network $\mathcal{R}$ that maps an utterance to a latent variable $\mathbf{z}$ and a generation network $\mathcal{G}$ that will be used to train $\mathbf{z}$'s representation. The goal is to represent the latent variable $\mathbf{z}$ independently of the context $\mathbf{c}$, so it can capture general dialogue semantics. LAED is a HRED model and the authors have introduced two versions of the model: Discrete Information Variational Auto-Encoder (DI-VAE) and Discrete Information Variational Skip-Thought (DI-VST).

DI-VAE works as a typical VAE by reconstructing the input $\mathbf{x}$ and minimizing the error between the generated and the original data. The loss function that optimizes the VAE model can be described as:

$$\begin{aligned}\mathcal{L}_{DI-VAE} = \mathbb{E}_{q_{\mathcal{R}}(\mathbf{z}|\mathbf{x})p(\mathbf{x})}[\log p_{\mathcal{G}}(\mathbf{x}|\mathbf{z})] \\ - KL(q(\mathbf{z}) \parallel p(\mathbf{z})),\end{aligned} \tag{2}$$

where $p(\mathbf{z})$ and $q(\mathbf{z})$ are, respectively, the prior and posterior distributions of $\mathbf{z}$, KL is the Kullback-Leibler divergence and $\mathbb{E}$ is the expectation.

DI-VAE model aims to capture utterance representations by reconstructing each word of the utterance. However, it is also possible to capture the meaning by inferring from the surrounding context, as dialogue meaning is very context-dependent. With this, the authors propose another version, the DI-VST, which is inspired by the Skip-Thought representation [17]. DI-VST uses the same recognition network from DI-VAE to output the posterior distribution $q(\mathbf{z})$, however, two generators are now used to predict both previous $\mathbf{x}_p$ and following $\mathbf{x}_n$ utterances. The loss function that optimizes DI-VST can now be described as:

$$\begin{aligned}\mathcal{L}_{DI-VST} = \mathbb{E}_{q_{\mathcal{R}}(\mathbf{z}|\mathbf{x})p(\mathbf{x})}[\log p^n_{\mathcal{G}}(\mathbf{x}_n|\mathbf{z}) \log p^p_{\mathcal{G}}(\mathbf{x}_p|\mathbf{z})] \\ - KL(q(\mathbf{z}) \parallel p(\mathbf{z})).\end{aligned} \tag{3}$$

DiKTNet learns this domain-agnostic representation from a large data-source and uses LAED models to perform this task. DiKTNet uses the DI-VAE model to obtain a latent representation of the user's request $\mathbf{z}_{usr}$ = DI-VAE($\mathbf{x}_{usr}$). As for the system's response, the model also wants to predict a latent representation $\mathbf{z}_{sys}$. In order to achieve that, DiKTNet uses the DI-VST model together with a context-aware hierarchical encoder-decoder that takes as input the user's request $\mathbf{x}_{usr}$ and the context $\mathbf{c}$. This encoder-decoder is different from the DI-VST for the reason that this new model, instead of predicting the previous and the following utterances, is interested in only predicting the following utterance that, in fact, is the system's response. The authors argue that DI-VAE captures the user utterance representation and that DI-VST predicts the system's action. When training with minimal data from the target domain, and after learning the latent representations $\mathbf{z}_{usr}$ and $\mathbf{z}_{sys}$, these variables are incorporated into the HRED $\mathcal{F}$ by an updated version of the loss function from equation 1:

$$\begin{aligned}\mathcal{L}_{HRED} = \mathbb{E}_{p(\mathbf{x}_{usr},c)p(\mathbf{z}_{usr},\mathbf{x}_{usr})p(\mathbf{z}_{sys}|\mathbf{x}_{usr},c)} \\ [\log p_{\mathcal{F}^d}(\mathbf{x}_{sys}|\{\mathcal{F}^e(\mathbf{c}, \mathbf{x}_{usr}), \mathbf{z}_{usr}, \mathbf{z}_{sys}\})],\end{aligned} \tag{4}$$

where $\{\ \}$ is the concatenation operator. With this, we ensure that the decoder is conditioned on the latent representations inferred in the pre-training phase and can now fine-tune in the target domain by taking into account that domain-agnostic representations. DiKTNet is also augmented with ELMo's [2] deep contextualized representations as word embeddings.

Instead of performing joint training as in original work, we first train the model with only source domains and then fine-tune it using a few example dialogues from the target domain. Below, we present how we enhanced our base model's performance using an improved training strategy.

4. Meta-learning

As we referenced in section 1, better training techniques improve the overall system's performance when adapting to new unseen domains using minimal data. In the following sections, we present our chosen meta-learning algorithm and describe how we adapted this algorithm into our base model.

4.1. Model-Agnostic Meta-Learning

In section 2, we described DAML [3] which incorporates the MAML [4] algorithm into the *sequicity* model. This optimization-based meta-learning technique aims to learn a good initialization for the model on source domains that can be efficiently adapted to target domains using minimum fine-tuning.

More formally, in each iteration of MAML, two batches of the training corpus are sampled from a source domain d: $\mathcal{D}^d_s$ and $\mathcal{D}^d_q$ which are named, respectively, the *source* and the *query* set. Instead of calculating the gradient step and updating the model, in each episode, low-resource fine-tuning is simulated: the model's parameters θ are preserved and for each domain d in source domains, new temporary parameters are calculated according to:

$$\theta^d = \theta - \beta \nabla_\theta \mathcal{L}(\theta, \mathcal{D}^d_s), \tag{5}$$

where β is the inner learning rate. We could update the model's original parameters with the sum of the losses from all source domains, however, we choose to update the parameters after each domain iteration as this method performs better as presented by [18].

After each episode, the model's parameters are updated using the temporary ones calculated in equation 5:

$$\theta = \theta - \alpha \nabla_\theta \mathcal{L}(\theta^d, \mathcal{D}_q^d), \tag{6}$$

where α is the outer learning rate. As our model incorporates both context and knowledge-base information for each dialogue and as MAML also consumes too much memory, we instead adopt a lightweight version of the MAML algorithm that we describe below.

4.2. Reptile

Reptile [5] algorithm is a first-order meta-learning algorithm where instead of sampling two source and query sets, $k > 1$ batches are retrieved for each domain $\mathcal{D}^d = (\mathcal{D}_1^d, ..., \mathcal{D}_k^d)$ and used to create the temporary model's parameters. The loss for the temporary model is calculated using Adam [19] optimizer according to:

$$\theta^d = \text{Adam}^k(\theta, \mathcal{D}^d, \beta), \tag{7}$$

where β is the inner learning rate and k is the number of updates in $\mathcal{D}^d$. After each episode, the model's original parameters are updated using the ones calculated in equation 7:

$$\theta = \theta + \alpha(\theta^d - \theta), \tag{8}$$

where α is the outer learning rate. Reptile is shown in [5] to produce equivalent or even better updates than MAML while consuming lower memory.

4.3. DATML

Our final model, DATML, is an adaptation of the architecture of DiKTNet with a modified training technique, while maintaining the strong representation learning. Instead of two training stages as in original work, we split joint training into source training and fine-tuning:

1. **Pre-training:** we maintain the first phase, where we learn the latent general representations for each turn using DI-VAE and DI-VST models.
2. **Source training:** in this phase, we exclude all data from the target domain and improve the training method by employing the Reptile meta-learning algorithm.
3. **Fine-tuning:** finally, we fine-tune the model using only a few example dialogues from the target domain.

5. Experiments

In this section, we describe how we evaluated both ZSDG and DiKTNet baselines and DATML. We also analyze and suggest possible limitations of our approach.

5.1. Datasets

The dataset used to obtain the latent actions in the pre-training phase for DiKTNet and DATML was the MetalWOZ dataset. MetalWOZ [20] is a dataset specifically constructed for the task of generalizing to unseen domains and is designed to help developing meta-learning models. This dataset contains about 37k task-oriented dialogues in 47 domains, such as schedules, apartment search, alarm setting, and banking. The data was collected in a Wizard-of-Oz fashion where a person acted like a robot/system and another acted as the user.

Table 1: *Excluded domains from MetalWOZ for each target domain on MultiWOZ dataset.*

MultiWOZ	**MetalWOZ**
hotel	*HOTEL_RESERVE*
restaurant	*MAKE_RESTAURANT_RESERVATIONS* *RESTAURANT_PICKER*
attraction	*EVENT_RESERVE*

Both baselines and our approach were evaluated on the three most represented domains from Multi-Domain Wizard-of-Oz dataset [7]: hotel, restaurant and attraction, where each contains more than 1500 dialogues. MultiWOZ is a large-scale multi-domain corpus containing human-to-human conversations with rich semantic labels (dialogue acts and domain-specific slot-values) from various domains and topics, and, like MetalWOZ, was collected in a Wizard-of-Oz fashion.

5.2. Experimental Setup

In the pre-training stage, we choose to learn the latent representations on MetalWOZ dataset as it is a domain-agnostic corpus introduced specifically for learning general representations. In order to make the evaluation as fair as possible, we exclude all dialogues from domains on MetalWOZ that could relate with the target domain on MultiWOZ, as described in table 1.

For source training, we train DATML on MultiWOZ dataset and exclude all dialogues from the target domains, including the multi-domain dialogues that contain turns from the target domain. In the fine-tuning phase, we use low resource data that varies from 1% to 10% by following [1] approach.

For both baselines and DATML, we follow [8] and [1] original setting and use Adam optimizer with a learning rate of 10^{-3} and Dropout ($p = 0.3$) [21]. All RNNs have hidden size of 512 and were trained for 50 epochs, using early stopping if the validation accuracy does not improve on half of already completed epochs. In the pre-training phase, we train both DI-VAE and DI-VST based LAED with y size of 10 and k size of 5, where y represents the number of latent variables and k the number of possible discrete values for each variable. For Reptile, we use a k size of 5 and train the model for 4000 episodes. The inner and outer learning rates are 10^{-3} and 10^{-1}, respectively.

For ZSDG, we followed the original author's [8] setting and used 150 seed responses for each domain. In order to fairly compare our model with state-of-the-art DiKTNet, we choose the same domain target data for both models by setting the random seed to 271, with no particular reason for selecting that number.

5.3. Metrics

We follow the work from DiKTNet [1] and ZSDG [8] and report BLEU and Entity F1 for each domain. These scores are calculated for each turn, where BLEU measures the similarity between the predicted and the reference responses and Entity F1 determines the ability of the model to retrieve correct entities from the knowledge base.

6. Results and Discussion

Table 2 shows results on MultiWOZ dataset. As observed in bold values, DATML outperforms both baselines ZSDG and DiKTNet in all low-resource scenarios.

Table 2: *Results on MultiWOZ dataset.*

Domain	hotel		restaurant		attraction	
Model	BLEU %	Entity F1 %	BLEU %	Entity F1 %	BLEU %	Entity F1 %
ZSDG	5.0	8.0	4.7	14.3	6.0	16.0
DiKTNet - 1%	10.7	17.3	12.4	17.5	10.2	18.6
DiKTNet - 3%	11.4	18.2	13.4	26.0	12.4	20.6
DiKTNet - 5%	11.6	17.6	16.6	25.7	12.0	27.1
DiKTNet - 10%	13.1	16.8	16.9	28.2	12.3	27.4
DATML - 1%	**10.9**	**18.0**	**14.1**	**24.0**	**11.0**	**23.4**
DATML - 3%	**13.0**	**23.1**	**16.7**	**28.4**	**14.1**	**28.6**
DATML - 5%	**14.1**	**25.3**	**17.8**	**30.0**	**15.0**	**31.2**
DATML - 10%	**14.2**	**26.3**	**18.3**	**32.9**	**15.4**	**32.2**

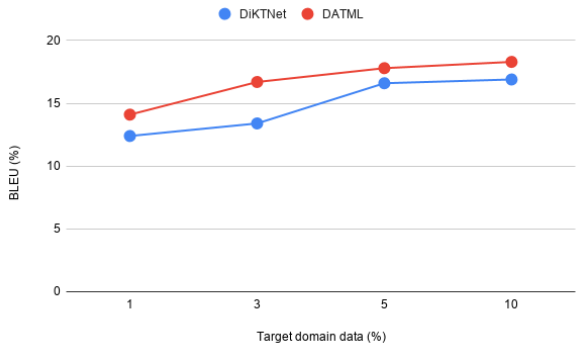


Figure 1: *BLEU score for different amounts of target data in the restaurant domain.*

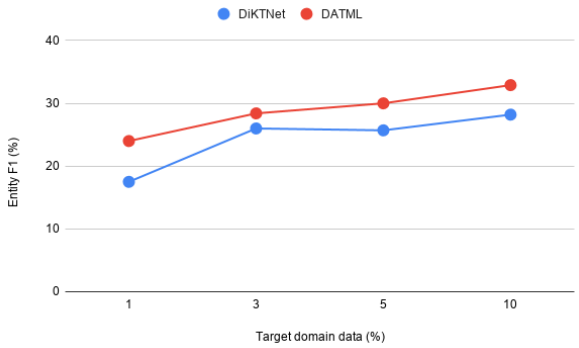


Figure 2: *Entity F1 score for different amounts of target data in the restaurant domain.*

We investigate how the use of different amounts of target domain data has an impact in the system's performance. Table 2 shows that our model's performance correlates with the amount of available data from the unseen domain. Figures 1 and 2 reveal that correlation for the restaurant domain and compare DiKTNet and DATML in terms of data usage. While small improvements can be observed when only 1% of target domain data is available, for each domain DATML achieves better results with 3% of target data in all metrics in comparison to DiKTNet with 10% of available target data. This shows that DATML outperforms DiKTNet in terms of both performance and data efficiency.

Table 2 also confirms that DiKTNet and DATML outperform ZSDG while using no annotated data and thus discarding human effort in annotating dialogues. This confirms that DATML achieves state-of-the-art results in data-efficiency and that is most suitable for real-world applications, as in underrepresented domains the amount of annotated data is almost nonexistent.

The results demonstrate that using optimization-based meta-learning improves the overall model's performance, and validate our initial idea that better learning techniques are a key feature when adapting to unseen domains using minimal data. Although our results seem promising and DATML outperforms previous state-of-the-art DiKTNet, these low scores are far from being sufficient for real-world applications, and more work is essential to surpass the problem of data scarcity in dialogue systems.

7. Conclusions

Domain adaptation in dialogue systems is extremely important as most domains are underrepresented. We proposed a model that improves previous state-of-the-art method by enhancing the training method. However, the evaluation results indicate that our model is far from being suited for real-world applications and show that this field requires more study. Future work includes improving the latent representations' retrieval and integration into our model. We would also like to refer that after submitting this paper we started some experiments with BERT-based [22] embeddings which are left for future work.

8. Acknowledgements

This work has been partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020.

9. References

[1] I. Shalyminov, S. Lee, A. Eshghi, and O. Lemon, "Data-efficient goal-oriented conversation with dialogue knowledge transfer networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1741–1751. [Online]. Available: https://www.aclweb.org/anthology/D19-1183

[2] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: https://www.aclweb.org/anthology/N18-1202

[3] K. Qian and Z. Yu, "Domain adaptive dialog generation via meta learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2639–2649. [Online]. Available: https://www.aclweb.org/anthology/P19-1253

[4] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 1126–1135. [Online]. Available: http://proceedings.mlr.press/v70/finn17a.html

[5] A. Nichol, J. Achiam, and J. Schulman, "On First-Order Meta-Learning Algorithms," *arXiv e-prints*, p. arXiv:1803.02999, Mar. 2018.

[6] Z.-Y. Dou, K. Yu, and A. Anastasopoulos, "Investigating meta-learning algorithms for low-resource natural language understanding tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1192–1197. [Online]. Available: https://www.aclweb.org/anthology/D19-1112

[7] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić, "MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 5016–5026. [Online]. Available: https://www.aclweb.org/anthology/D18-1547

[8] T. Zhao and M. Eskenazi, "Zero-shot dialog generation with cross-domain latent actions," in *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1–10. [Online]. Available: https://www.aclweb.org/anthology/W18-5001

[9] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 2208–2217. [Online]. Available: http://proceedings.mlr.press/v70/long17a.html

[10] D. George, H. Shen, and E. Huerta, "Deep transfer learning: A new deep learning glitch classification method for advanced ligo," *arXiv preprint arXiv:1706.07446*, 2017.

[11] Y. Yao and G. Doretto, "Boosting for transfer learning with multiple sources," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1855–1862, 2010.

[12] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4077–4087. [Online]. Available: http://papers.nips.cc/paper/6996-prototypical-networks-for-few-shot-learning.pdf

[13] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," *arXiv preprint arXiv:1711.04043*, 2017.

[14] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[15] W. Lei, X. Jin, M.-Y. Kan, Z. Ren, X. He, and D. Yin, "Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1437–1447. [Online]. Available: https://www.aclweb.org/anthology/P18-1133

[16] T. Zhao, K. Lee, and M. Eskenazi, "Unsupervised discrete sentence representation learning for interpretable neural dialog generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1098–1107. [Online]. Available: https://www.aclweb.org/anthology/P18-1101

[17] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015, pp. 3294–3302.

[18] A. Antoniou, H. Edwards, and A. J. Storkey, "How to train your MAML," *CoRR*, vol. abs/1810.09502, 2018. [Online]. Available: http://arxiv.org/abs/1810.09502

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] "A dataset of multi-domain dialogs for the fast adaptation of conversation models," https://www.microsoft.com/en-us/research/project/metalwoz/.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[22] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

AUTOMATIC SPEAKER ADAPTATION ASSESSMENT BASED ON OBJECTIVE MEASURES FOR VOICE BANKING DONORS

Agustin Alonso, Victor García, Inma Hernaez, Eva Navas, Jon Sanchez

HiTZ Basque Center for Language Technologies - Aholab, University of the Basque Country UPV/EHU

agustin@aholab.ehu.eus, {victor.garcia, inma.hernaez, eva.navas, jon.sanchez}@ehu.eus

Abstract

Speech is the most common way of communication. People who have lost total or partially their ability to speak might benefit from the use of Alternative and Augmentative Communication (AAC) devices and the use of Text-to-Speech (TTS) technology. One problem that arouses is that the synthetic voices included in these devices might be impersonal and not accurate to the user terms of age, accent or even gender. Therefore, voice banking has become a good alternative to standard commercial voices. In our voice banking strategy, people with healthy voice (donors), or the user itself before losing his or her own voice, provide the recordings to obtain a new synthetic voice using adaptation techniques. In this way, a wide catalog of synthetic voices is provided to the potential user. However, because there is no control over the recording process, the final quality of the synthetic voice is very variable. In this paper, we propose a method to assess the result of the adaptation using objective measures. The results show that this strategy can be an alternative to subjective evaluation to select the best donated voices for the voice bank.

Index Terms: STOI, ESTOI, NISQA, speech adaptation, voice banking

1. Introduction

Speech is the most common and intuitive way to communicate between humans. Sadly, in some cases people lose total or partially their ability to speak due to an accident or illness. In these cases, speech technologies can help those with some speech impairment. One possibility is providing them with a Text-to-Speech (TTS) system, so entering a text as input the system can reproduce the output with a synthetic voice. In general, this solution leads to general and non-emotive voices. Furthermore, the default voices could not suit the user preferences in terms of age, accent or gender (e.g. a young woman could use a voice of an old man). A solution is providing them a personalized voice. With a small amount of samples of a healthy voice, a new voice can be adapted for the TTS system. In this context voice banking has become a popular solution. If the person who wants to use the system still preserves his/her voice, he/she can make the necessary recordings before loosing his/her voice (due a scheduled surgery or a degenerative illness) and use his/her own adapted voice as a backup. Another approach consists on altruistic donors recording their voice which is used to train new ones for impaired people. Then, the user can choose their favourite one from the voice bank catalog. In these cases voice banks can handle hundreds or even thousands of donors, so an important task is classifying the result of adapting all these donors. The quality of the adapted voice, understood as how clean and intelligible it sounds, depends on several factors (initial quality of the recordings, how accurate the segmentation done during the training is, etc.) and measuring it manually is a great effort.

In this paper we analyze a method for automatically assess the adapted synthetic voices of our voice bank based on objective measures. We use two objective measures typically used in speech enhancement STOI [1] and ESTOI [2] and an objective measure based on NISQA that tries to estimate the naturalness MOS (Mean Opinion Score) of synthetic speech [3].

The rest of the article is structured as follows: in section 2 we explain voice banks in general and ours in more detail. Section 3 describes the objective measures we use in our analysis. The proposed analysis method is described in section 4 and the experiments performed are presented in section 5. Finally, in section 6 the main conclusions of this work are drawn.

2. Voice Banks

Voice banks are an alternative to provide people with speech difficulties with a personalized voice. In these voice banks, a personalized synthetic voice can be generated using adaptation techniques applied to a small sample of a healthy voice. There are several voice banks like ModelTalker [4], Speak Unique [5], VocalID [6] and the Voice Keeper [7].

Our voice bank, ZureTTS [8], offers the possibility of obtaining a personalized voice in Spanish and Basque. It makes use of a statistical synthesis engine based on Hidden Markov Models (HMMs) [9]. Each user must record a total of 100 phonetically balanced sentences in the selected language. These are parameterized using ahocoder [10], a high-quality vocoder that extracts MCEP coefficients of order 39, log-f_0 and maximum voice frequency. These data are then used to adapt an average voice using state of the art adaptation techniques [11] based on Constrained Maximum Likelihood Linear Regression plus Maximum a Posteriori Adaptation (CMLLR+MAP). For Spanish, the average voice was obtained with the subset 'phonetic' from the Albayzin [12] database. It consists of 6800 phrases from 204 different speakers in which each one has recorded 160, 50 or 25 sentences. For Basque, the average voice was obtained using the database described in [13]. This consists of 9 speakers (5 female and 4 male) all of which include 1 hour of speech except for two (female and male) which include 6 hours each. Currently, our voice bank has almost 9000 registered users.

3. Objective Measures Overview

3.1. STOI and ESTOI

In the field of intelligibility, several algorithms have been proposed that try to replace expensive subjective listening tests. Among them, STOI (Short Time Objective Intelligibility) [1] has proven to be good for evaluating intelligibility in signals

10.21437/IberSPEECH.2021-45

to which time-frequency weighting is applied. The method requires both the signal to be evaluated and a clean time-aligned reference. It calculates the Time-Frequency (TF) representation of both signals with a Discrete Fourier Transform (DFT) of the windowed frames and, using a one-third octave analysis, it groups the bins of the DFT into 15 bands and computes the norm of each one, which is called TF-unit. It uses an intermediate measure of intelligibility for each TF-unit, which depends on N consecutive TF-units of both the signal to be evaluated and the reference. Typically the value of N is that so the intermediate measure depends on speech information from the last $\approx 400ms$. To calculate the global intelligibility measure, the average of the intermediate intelligibility measurements between frames and frequency bands is computed. This operation implies an independent contribution to the global measure of each band. STOI has proven to predict intelligibility quite accurately in different situations, such as mobile phone output [14], noisy speech processed by ideal time-frequency masking and single channel speech enhancement algorithms [15] and speech processed by cochlear implants [16] and it is also robust against different types of languages like Mandarin [17], Danish [15] or Dutch [18].

One evolution of STOI is Extended STOI (ESTOI) [2], which unlike STOI does not assume independence between frequency bands. This feature allows to better capture the effect of time-modulated noise maskers.

The success of these measures has led to propose using them in several areas, for example, to evaluate the intelligibility of dysarthric speech [19]. In this work, STOI - ESTOI could not be applied directly since the time-aligned reference signal was not available. To overcome this problem, an utterance-dependent reference signal was generated from several healthy speakers and Dynamic Time Warping (DTW) was used to align the pathological signal and the reference signal.

3.2. NISQA

An important aspect to evaluate in synthetic voices is naturalness. In [3] a method based on Non-Intrusive Speech Quality Assessment (NISQA) [20] is proposed to measure the naturalness of synthetic voices without the need for a reference signal. The prediction model they propose is based on a CNN-LSTM network architecture with transfer learning domain knowledge from a speech quality database. It has been trained using 16 databases with 12 different languages, so it is presented as a language independent method that can be used in any TTS. Furthermore, the model has been trained using signals with different speech levels, to be able to be used with different types of signals. This makes it suitable for our voice banking case, were the adapted models are obtained with recordings without any control about speech level. The model is publicly available at [21] so it can be used directly.

4. Proposed Methodology

In this section we describe the method followed to obtain the objective measures. For STOI and ESTOI, the original recordings for each speaker are set as clean references and the synthetic signals are generated using the adapted voice model for each speaker. To obtain the NISQA score no processing of the synthetic signals is required. Once the predicted scores are obtained, we performed a clustering to identify the representative speakers, which will be included in the subjective evaluation.

4.1. STOI-ESTOI

The first step is to obtain the phonetic segmentation of the recordings. This is done by forced alignment using Montreal Forced Aligner (MFA) [22]. This step will be useful for two main reasons:

- It will provide the actual positions for the pauses made by the speaker during the recordings. The synthetic signals will be generated using this information, thus with the pauses at the same locations. We must consider that announcers can skip pauses indicated by spelling signs, or on the contrary, make pauses that are not in the text.
- Have the segmentation available for later use in the alignment between the reference and the signal whose intelligibility is going to be evaluated.

Next, the synthetic signals corresponding to each speaker are generated, with the previously detected pauses. In this way, a parallel corpus of recordings-synthetic signals is available. However, these signals will have different duration, and therefore it will be necessary to align them. To do this, although they could be aligned at sentence level, in our system we perform an alignment with DTW at phoneme level. For the alignment, the cepstral distances between reference and 'target' are calculated. The cepstral coefficients are obtained directly from the adapted voice model for synthetic signals, while for reference signals, they are obtained using Ahocoder [10]. It has also been necessary to adapt the frame rate of the synthesis system to the one used by the STOI / ESTOI algorithm.

After these steps, a score can be obtained with the STOI and ESTOI algorithm for each sentence. The final score for each speaker will be the average of the scores obtained for the 100 available sentences.

4.2. NISQA

As this measurement does not require any reference, to calculate the donor's NISQA score, the score of all the previously generated synthetic sentences is calculated and averaged.

4.3. Clustering

We have three objective measures for each donor: STOI, ESTOI and NISQA. Since the scale of the three is different, the z-score of each one is calculated by normalizing by mean and variance so that all of them have the same importance in the clustering process. The donors are then grouped by applying the k-means clustering algorithm using the normalized objective measures.

5. Experiments

5.1. Experimental Setup

The proposed method has been tested using the recordings and the adapted synthetic voices in Spanish from ZureTTS voice bank [23]. A total of 1090 voices have been used.

Considering that users evaluate the quality of the synthetic voices using a 5 grade scale, we have arbitrarily set a final number of 5 clusters. In order to find out which objective measure is more important for the users when evaluating their preference for one synthetic voice or another, the clustering has been done in two different ways.

A) Taking into account only intelligibility related measures, i.e. STOI and ESTOI

B) Taking into account the three measures

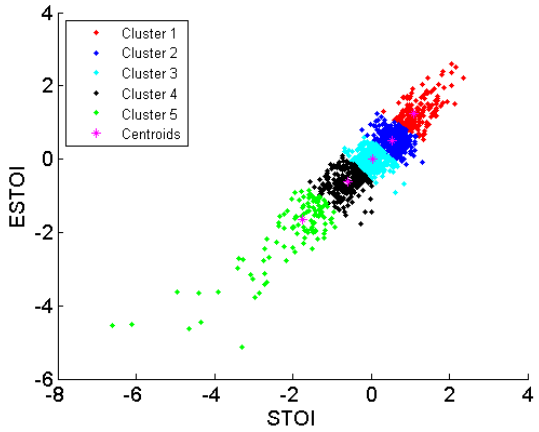


Figure 1: *Clustering using normalized STOI and ESTOI averaged scores.*

The scatter plots of figures 1 and 2 show the computed normalized scores and the resulting clusters. Figure 1 corresponds to the clustering performed when only STOI and ESTOI measures are used. As expected both measures are highly correlated ($\rho = 0.912$) and the clustering shows clear linear boundaries. When the naturalness score is incorporated (Figure 2) the final clusters are modified. The changes mainly affect to clusters 2 and 3, which are separated in the NISQA dimension (figures 2b and 2c), but mixed in the STOI and ESTOI dimensions (figure 2a).

As reference donors to represent each cluster, the centroids are chosen. The objective measures values for these centroids are shown in table 1. Only the measures used for the clustering are shown.

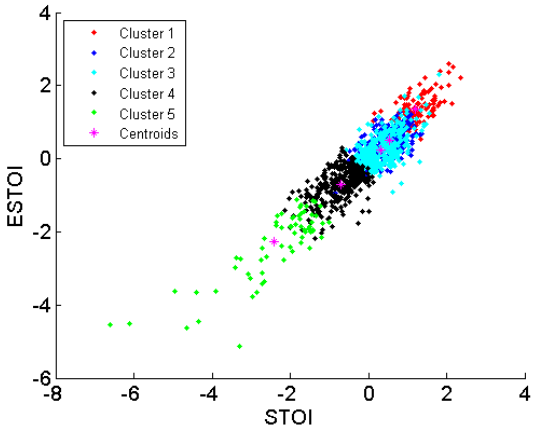
Table 1: *Objective measures for representative donors*

Donor	STOI	ESTOI	NISQA
SPK1	0.6832	0.5045	-
SPK2	0.6689	0.4689	-
SPK3	0.6155	0.4220	-
SPK4	0.5747	0.3790	-
SPK5	0.5015	0.3093	-
SPK6	0.6895	0.5122	3.0036
SPK2	0.6689	0.4689	2.8210
SPK7	0.6343	0.4368	2.5261
SPK8	0.5691	0.3732	2.4926
SPK9	0.4601	0.2679	1.1778

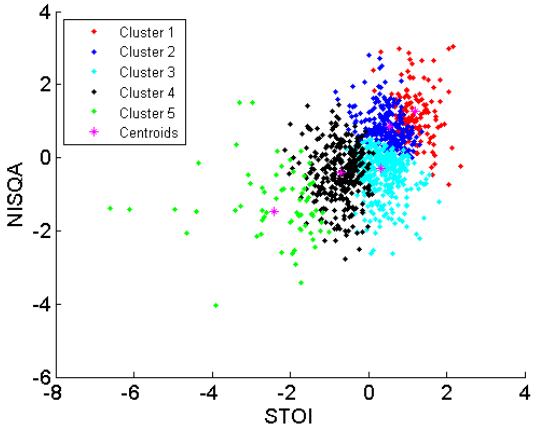
Speakers SP1 to SPK5 are the centroids corresponding to clustering A, while speakers SPK6 to SPK9 correspond to the centroids of clustering B. SPK2 coincided as a centroid in both clusterings.

5.2. MOS Evaluation

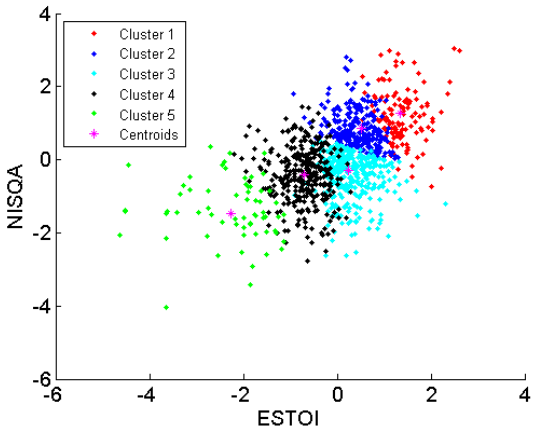
To check how the clustering results agree with people's preferences, a subjective evaluation was carried out. Using the adapted models from the 9 representative donors 10 new short sentences were synthesized, i.e. 90 sentences in total. 15 people took part in the evaluation, 5 of them were experts in speech technologies. Each one scored 5 randomly selected sentences



(a) *Clustering projection over the STOI-ESTOI axis.*



(b) *Clustering projection over the STOI-NISQA axis.*



(c) *Clustering projection over the ESTOI-NISQA axis.*

Figure 2: *Clustering using normalized STOI, ESTOI and NISQA averaged scores.*

from each donor, so each evaluator evaluated 45 sentences. The sentences were presented in a simple web interface where for each case they had to score on a MOS scale of 1 to 5 a single question: "Would you use this synthetic voice in a TTS system?" where 1 meant "No, I do not like it and I would not use

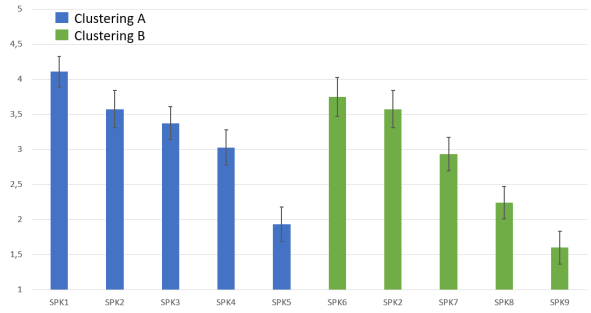


Figure 3: *MOS results with 95% confidence interval.*

it" and 5 meant "Yes, I like how it sounds and I would to use it". Some evaluation samples can be listened here[1].

The results of the subjective evaluation are shown in figure 3.

As it can be observed in the figure, the ordering of the speakers is the same for the two considered classification systems (Clustering A and Clustering B). In fact, the three objective measures, considered independently, are highly correlated with the obtained subjective MOS scores. Table 2 shows the correlation coefficient between the obtained final scores, using the nine donors for STOI and ESTOI and the five donors of Clustering B for NISQA. We can see that intelligibility scores correlation with MOS is higher than for naturalness. We can also observe from figure 3 that the set of donors from clustering A have obtained better scores than those from Clustering B. A possible interpretation is that when choosing a synthetic voice, the features measured by the NISQA algorithm are not as relevant as those measured by the intelligibility measures STOI and ESTOI for the evaluators. Also, naturalness is not relevant when intelligibility is not guaranteed, as is the case for some adapted voices in our system.

Table 2: *Correlation coefficient between subjective MOS and objective measures*

	STOI	**ESTOI**	**NISQA**
ρ	0,9484	0,9500	0,7858

6. Conclusions and Future Work

We have seen how the use of objective measures on the result of the adaptation in our voice bank can be used to make a first classification in the voice catalog. The subjective evaluation using the representative donors of each cluster confirms that the better the objective measures, the better the acceptance of the synthetic voice by the users. We can use the result of the clustering to do an initial categorization and set the donors from the first cluster as voice bank catalog. Clustering performed without using NISQA naturalness measure has resulted in more MOS correlated scores, so calculation of only STOI and ESTOI seems enough to estimated people's perception. However, as future work, we plan to consider more measures such as DAU [24] or GP (Glimpse Proportion) [25] [26] to the study, trying to improve clustering and thus a more accurate automatic classification of adapted voices.

[1] https://aholab.ehu.eus/users/agustin/demos/ib20/

7. Acknowledgements

This work has been funded by the Basque Government under the project ref. PIBA 2018-035 and IT-1355-19.

8. References

[1] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[2] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[3] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," *Proc. Interspeech 2020*, pp. 1748–1752, 2020.

[4] *Model Talker*. [Online]. Available: https://www.modeltalker.org/

[5] *Speak Unique*. [Online]. Available: https://www.speakunique.co.uk/

[6] *VocalID*. [Online]. Available: https://vocalid.ai/

[7] *The Voice Keeper*. [Online]. Available: https://thevoicekeeper.com/

[8] D. Erro, I. Hernáez, E. Navas, A. Alonso, H. Arzelus, I. Jauk, N. Q. Hy, C. Magariños, R. Pérez-Ramón, M. Sulír, X. Tian, X. Wang, and J. Ye, "ZureTTS: Online platform for obtaining personalized synthetic voices," in *Proc. eNTERFACE'14*, 2014.

[9] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[10] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus Noise Model based Vocoder for Statistical Parametric Speech Synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 184–194, 2014.

[11] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Audio, Speech, & Lang. Process.*, vol. 17, no. 6, pp. 1208–1230, 2009.

[12] M. A. Moreno Bilbao, D. Poig, A. Bonafonte Cávez, E. Lleida, J. Llisterri, J. B. Mariño Acebal, and C. Nadeu Camprubí, "Albayzin speech database: Design of the phonetic corpus," in *EUROSPEECH 1993: 3rd European Conference on Speech Communication and Technology: Berlin, Germany: September 22-25, 1993*. . EUROSPEECH, 1993, pp. 175–178.

[13] I. Sainz, D. Erro, E. Navas, I. Hernáez, J. Sanchez, I. Saratxaga, and I. Odriozola, "Versatile speech databases for high quality synthesis for basque." in *LREC*. Citeseer, 2012, pp. 3308–3312.

[14] S. Jørgensen, J. Cubick, and T. Dau, "Speech intelligibility evaluation for mobile phones," *Acta Acustica United with Acustica*, vol. 101, no. 5, pp. 1016–1025, 2015.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[16] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE signal processing magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[17] R. Xia, J. Li, M. Akagi, and Y. Yan, "Evaluation of objective intelligibility prediction measures for noise-reduced signals in mandarin," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4465–4468.

[18] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 2, pp. 430–440, 2014.

[19] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Pathological speech intelligibility assessment based on the short-time objective intelligibility measure," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6405–6409.

[20] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7125–7129.

[21] *NISQA*. [Online]. Available: https://github.com/gabrielmittag/NISQA

[22] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: trainable text-speech alignment using kaldi," in *Proceedings of INTERSPEECH*, 2017, pp. 498–502.

[23] D. Erro, I. Hernaez, A. Alonso, D. García-Lorenzo, E. Navas, J. Ye, H. Arzelus, I. Jauk, N. Q. Hy, C. Magariños *et al.*, "Personalized synthetic voices for speaking impaired: Website and app," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[24] C. Christiansen, M. S. Pedersen, and T. Dau, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Communication*, vol. 52, no. 7-8, pp. 678–692, 2010.

[25] M. Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.

[26] Y. Tang, M. Cooke *et al.*, "Glimpse-based metrics for predicting speech intelligibility in additive noise conditions." in *Interspeech*, 2016, pp. 2488–2492.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Data-driven analysis of nasal vowels dynamics and coordination in bilabial contexts

Conceição Cunha[*2], *Nuno Almeida*[*1], *Jens Frahm*[3], *Samuel Silva*[1], *António Teixeira*[1]

[*]These authors contributed equally to this work
[1]IEETA, DETI, University of Aveiro, Aveiro, Portugal
[2]Institute of Phonetics and Speech Processing, LMU Munich, Germany
[3]Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany
cunha@phonetik.uni-muenchen.de, nunoalmeida@ua.pt

Abstract

One of Portuguese distinctive marks is the large nasals inventory, including five phonemic nasal vowels and three diphthongs. Previous studies argued for an initial oral part and a short nasal consonant, probably related to synchronization between oral and nasal gestures. These studies have considered discrete descriptions with EMA-flesh points, limiting our grasp of the whole vocal tract, and preliminary work using real-time MRI (RT-MRI) considered a small frame rate (14fps) and a reduced number of speakers, yielding a rather small time-resolution to study an intrinsically dynamic process. The recent advances of RT-MRI, with frame rates of 50fps, have made possible a finer detail of the dynamics of nasals. However, new challenges need to be tackled to deal with the resulting large amount of data and to foster analyses to tackle a larger number of speakers. Grounded on a new RT-MRI corpus for European Portuguese, this paper explores the capabilities of recent data-driven methods, to analyze dynamic aspects of nasal vowels and coordination. To this end, we consider data for 11 EP speakers and investigate vocal tract configurations, over time, and the coordination of velum and lip aperture in bilabial (oral and nasal) contexts. Overall, the results show changes of the vocal track and the model explores the dynamic behavior of the vowel tract along the production of oral and nasal vowels.
Index Terms: Speech production studies, European Portuguese, RT-MRI, nasals, data-driven analysis

1. Introduction

European Portuguese (EP) distinguishes five nasal vowels and three diphthongs. Nasal vowels are complex sounds visible also in the acoustic antiformants caused by the inclusion of the nasal cavity in the production and the difficult parsing of articulation and acoustics. EP nasal vowels are usually divided in an initial oral portion, a nasal part and a consonantal tail, probably related to a late alignment of the velum relatively to the vowel tract configuration for the vowel production. However, there is no robust evidence for this partition and this will be one of the advances of this paper. These dynamic aspects cannot be caught with static analysis of vowel midpoints and do have to consider the whole or a greater part of its duration.

The study of such complex sounds, exhibiting characteristic dynamic patterns contributes to improve our knowledge on speech production supporting a range of applications from speech therapy to articulatory speech synthesis. Initial studies of EP nasal vowels production dynamics resorted to Electromagnetic Articulography (EMA) data, and contemplated: quantitative analysis of velum movement in context of stop consonants [1]; comparison with French nasal vowels [2]; study of gestures timing, characterization of the gesture in terms of average duration, investigation of factors influencing such durations, and characterization of inter-gestural coordination [3]; and speech rate effects [4]. With advances in magnetic resonance imaging (MRI) and real time magnetic resonance imaging (RT-MRI), this studies were complemented by exploring the coverage of the complete vocal tract in, for example [5, 6, 7, 8].

It is important to expand the body of knowledge for EP nasals by moving into data acquisition techniques that provide a more complete view over the vocal tract, when compared to EMA, along with a finer grasp of tract dynamics. Advances in RT-MRI of the vocal tract [9] allow improving on previous research by providing better image quality at a higher frame rate of 50fps, enabling to move beyond past limitations on time resolution. However, with this new data, several challenges arise mostly resulting from the sheer amount of data and from how to systematically process and analyze it to obtain useful results to address the research questions in a quantitative manner [10, 11]. Additionally, the amount of available data also opens new possibilities to move into more data-driven analysis, e.g., adopting methods to model the behavior of relevant sounds by gathering the data from multiple repetitions and/or speakers.

In view of recent methods proposed to process and analyze speech production data extracted from real-time MRI of the vocal tract [12] it is important to understand how these, proposed for a general scope, might be of use to investigate the dynamics of nasals. In this context, the work presented here considers a novel RT-MRI database for European Portuguese and aims to:

- Explore recent methods proposed to support dynamic analysis and modeling from articulatory data extracted from RT-MRI of the vocal tract;
- Assess the applicability of these new methods to the analysis of nasal vowels;
- Provide first reports regarding vowel dynamics and coordination for EP nasal cardinal vowels in bilabial contexts.

The remainder of this document is organized as follows. Section 2 presents an overall overview and background regarding the study of nasal vowels and the consideration of data-driven methods to process and analyze speech production data; section 3 describes the considered data and provides an overview of the methods to tackle it; section 4 presents and discusses the obtained results; and, finally, section 5 presents some conclusions and routes for further work.

10.21437/IberSPEECH.2021-46

2. Background and Related Work

2.1. On nasal vowels

The production of nasal vowels involve much more than lowering the velum. The way this aperture, and other articulators, vary in time and their coordination is important [13]. Nasal vowels can be regarded as diphthongs [14], starting with dominant lips radiation and ending in a nasal radiation dominant configuration. This dynamic nature of nasal vowels is important for their perception [15].

It is often unclear when a vowel is a phonemic nasal or simply contextually nasalized. For example, for Brazilian Portuguese, Meireles [16] found a synchronous coordination of the nasal vowel with the preceding consonant, while Desmeules-Trudel [17] reported a very late alignment of nasal and oral gestures, arguing against the phonemic status of the nasal vowels.

Despite efforts such as [3, 8, 10], it is not yet completely clear how oral and nasal gestures are synchronized in EP nasal vowels production, mainly due to the limitations of the analyzed data (restricted number of speakers and partial information regarding the tract, for EMA, or reduced temporal resolution, for RT-MRI).

2.2. Analysis and Modeling of Articulatory Data

With the increase of the data available from different technologies supporting speech production studies, as is the case for RT-MRI [18, 9], it is paramount to pursue methods that enable its systematic quantitative assessment through unsupervised approaches, to take the most out of the available data. In this regard, several authors have proposed data-driven methods for their processing and analysis (e.g. [19, 20, 21].

In this regard, the authors have explored data-driven approaches to determine critical articulators from vocal tract data extracted from RT-MRI. [22, 11] expanding an approach proposed for EMA [23]. One notable aspect of the presented approach is that, even though the method considers statistical modeling for the different sounds, the consideration of tract variables aligned with the Articulatory Phonology framework [24], as grounds for the method, yields results that keep a connection to the tract anatomy (e.g., constrictions on the tongue tip and body) and are, hence, more interpretable towards a critical discussion of the outcomes and an improved knowledge of speech production. Nevertheless, while this provides valuable information for a wide range of sounds, in an unsupervised manner, in its current state it still only considers a static representation for each phone (i.e., one time point along the production). Even though, for nasal vowels, three time points were selected, to grasp some of the dynamics, the granularity of the resulting data does not enable taking conclusions about the subtlety of the underlying dynamics and coordination.

In a recent article, Carignan et al. [12] explore how vocal tract data extracted from RT-MRI can be explored by modeling the dynamics of speech production based on multiple repetitions of each sound. In their method they adopt generalized additive mixed models (GAMMs) applied to vocal tract aperture functions, along with validations of the resulting models using functional linear mixed models (FLMMs) at 20% and 80% of the vowel interval. Overall, these methods model vocal tract aperture, over time, for given sounds, and can be useful to gain insight over how sounds are produced considering data from multiple speakers, at once. One notable point addressed is the interpretability of the results obtained with the proposed methods. To this end, the authors establish a correspondence between data points on the tract aperture functions and anatomical regions and apply this principle to all speakers and repetitions, which then allows an understanding of the resulting GAMMs.

3. Methods

In what follows, an overall description of the considered data and methods adopted for analysis is provided.

3.1. Data Aquisition

The RT-MRI dataset recordings were performed at the Max-Planck-Institute in Göttingen, Germany, using a 3T Siemens Prisma Fit MRI System equipped with a 64-channel head coil. The MRI acquisitions involved a low-flip angle gradient-echo sequence with radial encodings and a high degree of data under sampling [9]. The procedure allowed for real-time image sequences of the vocal tract in a midsagittal plane of the speaker at 50 fps. Speech sound was synchronously recorded using an optical microphone (Dual Channel-FOMRI, Optoacoustics) and annotated using Praat [25].

3.2. Corpus and Speakers

The analysed corpus consists of minimal pairs containing the three stressed oral and nasal point vowels [i, u, a] and [ĩ, ũ ɐ̃] preceded by bilabial oral or nasal consonants, as in the following words: 'pato' [patu], 'panto' [pɐ̃tu], 'mato' [matu], 'manto' [mɐ̃tu]. All words were randomized and repeated in two prosodic conditions embedded in one of three carrier sentences alternating the verb as follows: (diga ('Say'); ouvi ('I heard'); leio ('I read')) as in 'Diga pato, diga pato baixinho' ('Say duck, Say duck gently'). The data considered for the analysis presented in this article include 11 native speakers of EP and is part of a larger corpus being acquired to study EP nasals.

3.3. MRI data Processing and Analysis

Overall, our purpose was to explore the applicability of the method proposed by Carignan et al. [12]. In short, the processing pipeline consists of five steps: (1) **Image registration**, in which, images are aligned to compensate some movement of the speaker; a (2) **Semi-polar grid** is placed throughout the vocal tract, it consists of 28 lines distributed from the glottis to the anterior edge of the alveolar ridge; (3) **Air-tissue boundary detection** processes semi-automatically each frame to find the outer boundary of the vocal tract for each grid line; the (4) **Aperture estimation** is obtained by counting the pixels from the boundary that are below a determined threshold; (5) Principal Component Analysis (PCA) of the **velum and lip aperture** are estimated by an approach based on region-of-interests (ROI) and considering pixel intensities (e.g., the amount of darker pixels is higher for the inter-lip region when the lips are open).

Following the same approach as Carignan et al. [12], but since we are interested in lip and velar aperture, we have complemented the semi-polar grid data with the data obtained from the velum and lips to test if Generalized Additive Mixed Models (GAMMs) visualizations could provide any useful insight on their changes, over time. To this end, 1) we have added two more grid lines in the lips area with the computed values; and 2) in the velum area, we have subtracted a factor of the value obtained in the velum PCA from the tract aperture function, in that region, in an attempt that velar opening would induce a discernible change, in the GAMMs, between oral and nasal sounds, our object of study. To get a grasp of what is happening

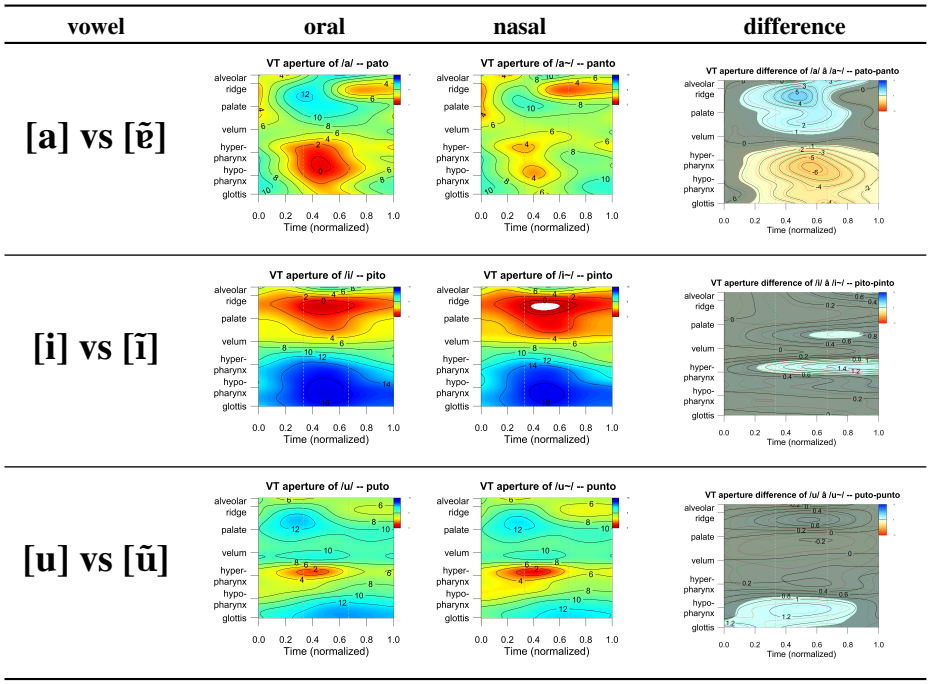


Figure 1: *Results from GAMMs for the three vowel pairs in /p/ context. From left to right, in each row, the plots refer to the oral, the nasal, Each GAMMs plot encompasses the previous consonant /p/, the vowel (at the center, from 0.33 to 0.66), and the next consonant. The results for the /m/ context presented similar patterns and are not shown for the sake of brevity.*

to the vocal tract, we also added the preceding and following consonant, as context. The obtained GAMMs show the vowel (at their center, between 0.33 and 0.66), but also show the preceding consonant (from 0 to 0.33) – a /p/ or a /m/, and the following consonant (from 0.66 to 1) – a /t/ or a /d/. See Figure 1 for examples.

To complement the GAMMs visualizations with more detailed data for analysis, Functional Linear Mixed Models (FLMMs) were computed for tract apertures at 20%, 50% and 80% of the vowel interval. Also, a first attempt is presented of using FLMMs to model the behaviour of the lips and velum, over the time, to support the analysis of their coordination.

4. Results

This section presents examples of analyses aiming at both assessing the capabilities of the methods and contributing to augment knowledge regarding the temporal aspects of nasal vowel production. The results gathered for 11 EP speakers covering an overall analysis of tract dynamics, a more detailed analysis of tract configuration for key timepoints along the vowels, and an overall analysis of lip and velar coordination.

4.1. Overall Vocal Tract Dynamics

The analysis started by an overall assessment of how the tract evolves for oral and nasal vowels by applying GAMMs to the considered bilabial contexts for the 3 vowels. The resulting 'areograms' showing data for the vowel, at the center of each plot (from 0.33 to 0.66), along with the flanking consonants ([p] and [t]) are presented in Figure 1. Each row shows the "areogram" for the oral, its nasal congener, and the difference between the two. For the sake of brevity and given no major differences between both bilabial contexts, only the oral [p] contexts are shown.

Overall, the differences for the pairs [u]/[ũ] and [i]/[ĩ] are very small, for both contexts, noticeable by the dominant grayish color of the difference diagram. Differences between [a] and [ẽ] are more noticeable. For both contexts, the oral is more backed than the nasal (the redish area on the difference GAMMs, for the lower tract, and blueish around the alveolar ridge). Finally, the dynamic pattern for each vowel pair is similar across the considered contexts. The GAMMs representations, obtained based on the tract area functions, which we modified by including velar and labial aperture, provide good insight on oral configurations, but nasality differences do not arise, in the representation, as we attempted.

4.2. Detailed Analysis at Specific Timepoints

While the GAMMs in Figure 1 show the overall evolution of tract aperture, to have a more detailed grasp of the tract's configuration, Figure 2 shows superimposed plots of tract apertures at specific times: beginning (20%), middle (50%), and end (80%) of the vowel interval.

For [a] and [ẽ] these plots further confirm the previously described results. For both contexts, the oral vowel exhibits a smaller aperture at the back of the tract pointing to a more backed configuration than the nasal. At the alveolar ridge, it is the opposite effect, consistent with the observed backness of the oral and hinting on a wider lip aperture (although the plots do not explicitly include lip data), and also, possibly, of the jaw.

For vowels [i] and [u], the differences towards their nasal congeners are very small (slightly higher for [u]) and no notable difference appears across contexts. For [u], the nasal shows a slight difference for the alveolar ridge, from 50% onward, possibly due to the tongue movement to produce the [n], in the considered contexts.

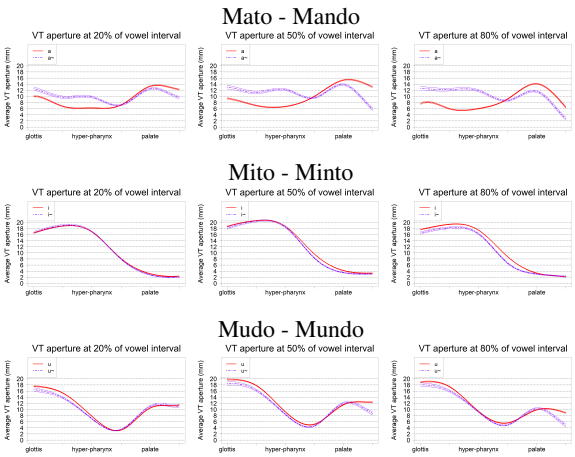


Figure 2: *FLMMs showing the aperture function of the vocal tract for three time points. Each row presents the aperture function for 20%, 50% and 80%, along the vowel. Each of the presented plots shows, in red, the curve for the oral and, in blue, the curve for the nasal. The curves for the oral bilabial context [p] showed similar patterns.*

4.3. Coordination

Since the vowels were produced in bilabial context, we will analyse the coordination of the lower lip with the velum. Figure 3 shows FLMMs for the lip and velar aperture, over time, for the different oral/nasal pairs. By modeling lip and velar behavior considering the data for the 11 speakers, we obtain a first grasp over coordination, an important (and challenging) aspect to study for nasal vowels. For vowel [ɐ̃], in both contexts, lip aperture's peak occurs earlier than for the oral vowel and to a smaller aperture. The smaller aperture is consistent with the more closed vowel quality of the nasal congener. Vowel [i] seems to show a similar (albeit less pronounced) pattern. For the nasals in [p] context, the onset of velar opening seems to happen slightly after lip occlusion release.

Overall, velar behavior for both contexts is as expected. For [p] contexts, the velum starts closed, opens during the nasal vowel, and closes, again. For [m], the velum starts open and gradually closes during the oral vowel, or continues closing after the nasal vowel. Interesting to note is the higher error interval for the velar curve in [ũ]. For ɐ̃, and ĩ, a close observation of the curves seems to reveal a slight tendency of lip closure minima, after the vowel, occurring slightly before velar closure.

5. Conclusions

This paper presents a novel analysis of EP nasal vowels regarding dynamics of the tract configuration and coordination of velum movement with lip aperture based in the application of GAMMs to high frame rate RT-MRI data.

The methods explored in this article enabled an elegant approach to tackling data from multiple speakers to reach an overall model of the dynamic behavior of the tract, along the production of oral and nasal vowels. The results presented, although a first exploration of these methods, already highlight interesting aspects regarding nasal vowels for a considerable number of speakers. These results corroborate the similarity of the vocal tract configuration, over time, for [ĩ] and [ũ], when compared with their oral congeners and a stronger difference in configuration between [a] and [ɐ̃] that seems to be slightly more pronounced at the end of the vowel.

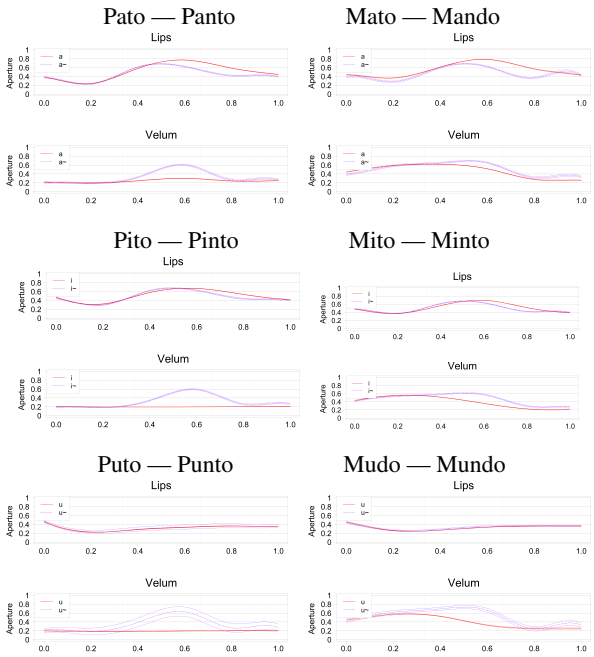


Figure 3: *FLMMs of lip and velar aperture, over time, for the different vowels and contexts, obtained from production data from 11 speakers. Curves for the oral vowels in red, and for the nasals in blue.*

A very interesting result is the stronger variation for the velar aperture for [ũ]. This is probably due to the backness of the vowel and the proximity of the tongue with the velum, potentially originating adjustments or less well defined timings (along with possible image artefacts, introducing a bit more uncertainty). This demands further investigation to disentangle the different possible causes. In this regard, also having data concerning the tongue body movement towards the pharynx might shed some light on what adjustments are involved. Regarding coordination, the FLMMs seem to provide a hint of an anticipatory closure/minimal area of the lips in respect to velar closure. However, to ascertain if this is confirmed – which would provide evidence supporting the occurrence of a small consonantic nasal tail – a more detailed analysis is required.

The changes introduced to the data input for the GAMMs, including data for the lips and velar aperture did not yield a clear depiction of the nasality differences among oral and nasal vowels. This might be a result of the smoothing effect of the GAMMs representation, given the limited spatial scope of the change. Further work to improve these aspects is required. In this regard, it would now be interesting to further test these methods with data extracted from tract segmentations to consider direct measures of the velopharyngeal passage and interlip distance as we considered in [11], also for RT-MRI.

6. Acknowledgements

A word of thanks is due to Dr Christopher Carignan for sharing the scripts serving as basis for the methods explored here. This work is partially funded by the German Federal Ministry of Education and Research (BMBF, KZ:01UL1712X), by IEETA Research Unit funding (UIDB/00127/2020), by Portugal 2020 under COMPETE Program, and the European Regional Development Fund through project SOCA – Smart Open Campus (CENTRO-01-0145-FEDER-000010), and project MEMNON (POCI-01-0145-FEDER-028976).

7. References

[1] A. Teixeira and F. Vaz, "European Portuguese nasal vowels: An EMMA study," in *7th European Conference on Speech Communication and Technology, EuroSpeech - Scandinavia*, vol. 2. Aalborg, Dinamarca: CPK/ISCA, Sep. 2001, pp. 1843–1846.

[2] S. Rossato, A. Teixeira, and L. Ferreira, "Les nasales du Portugais et du Français : une étude comparative sur les données EMMA," in *JEP'2006*, Rennes, França, 2006.

[3] C. Oliveira and A. Teixeira, "On gestures timing in European Portuguese," in *ICPhS*, 2007, pp. p. 405 – 408.

[4] C. Oliveira, P. Martins, and A. Teixeira, "Speech rate effects on European Portuguese nasal vowels," in *InterSpeech*, 2009.

[5] P. Martins, I. Carbone, A. Silva, and A. Teixeira, "An MRI study of European Portuguese nasals," in *Interspeech*, 2007.

[6] ——, "European Portuguese MRI based speech production studies," *Speech Communication*, vol. 50, pp. 925–952, 2008.

[7] A. Teixeira, P. Martins, C. Oliveira, C. Ferreira, A. Silva, and R. Shosted, "Real-time MRI for Portuguese," in *Computational Processing of the Portuguese Language, PROPOR 2012, Lecture Notes in Computer Science/LNAI, Vol. 7243*, 2012.

[8] C. Oliveira, P. Martins, S. Silva, and A. Teixeira, "An MRI study of the oral articulation of European Portuguese nasal vowels," in *13th Annual Conference of the International Speech Communication Association (InterSpeech)*, Portland, USA, September 2012.

[9] M. Uecker, S. Zhang, D. Voit, A. Karaus, K.-D. Merboldt, and J. Frahm, "Real-time mri at a resolution of 20 ms," *NMR in Biomedicine*, vol. 23, no. 8, pp. 986–994, 2010.

[10] C. Cunha, S. Silva, A. Teixeira, C. Oliveira, P. Martins, A. A. Joseph, and J. Frahm, "On the Role of Oral Configurations in European Portuguese Nasal Vowels," in *Proc. Interspeech 2019*, 2019, pp. 3332–3336. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2232

[11] S. Silva, N. Almeida, C. Cunha, A. Joseph, J. Frahm, and A. Teixeira, "Data-driven critical tract variable determination for european portuguese," *Information*, vol. 11, no. 10, p. 491, 2020.

[12] C. Carignan, P. Hoole, E. Kunay, M. Pouplier, A. Joseph, D. Voit, J. Frahm, and J. Harrington, "Analyzing speech in both time and space: Generalized additive mixed models can uncover systematic patterns of variation in vocal tract shape in real-time MRI," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 11, no. 1, 2020.

[13] A. Teixeira, F. Vaz, and J. C. Príncipe, "Nasal vowels after nasal consonants," in *5th Seminar on Speech Production: Models and Data*, Kloster Seon, Alemanha, May 2000.

[14] S. Parkinson, "Portuguese nasal vowels as phonological diphthongs," *Lingua*, vol. 61, no. 2-3, pp. 157–177, 1983.

[15] A. Teixeira, F. Vaz, and J. C. Príncipe, "Influence of Dynamics in the Perceived Naturalness of Portuguese Nasal Vowels," in *14th International Congress of Phonetic Sciences (ICPhS 99)*, San Francisco, CA, E. U. A., Agosto 1999.

[16] A. R. Meireles, L. Goldstein, R. Blaylock, and S. S. Narayanan, "Gestural coordination of brazilian portugese nasal vowels in cv syllables: A real-time mri study." in *ICPhS*, 2015.

[17] F. Desmeules-Trudel, "The aerodynamics of vowel nasality and nasalization in brazilian portuguese." in *ICPhS*, 2015.

[18] A. D. Scott, M. Wylezinska, M. J. Birch, and M. E. Miquel, "Speech MRI: Morphology and function," *Physica Medica*, vol. 30, no. 6, pp. 604 – 618, 2014.

[19] A. C. Lammert, M. I. Proctor, S. S. Narayanan *et al.*, "Data-driven analysis of realtime vocal tract MRI using correlated image regions." in *Proc. Interspeech*, 2010, pp. 1572–1575.

[20] Q. Chao, "Data-driven approaches to articulatory speech processing," Ph.D. dissertation, University of California, Merced, 2011.

[21] M. P. Black, D. Bone, Z. I. Skordilis, R. Gupta, W. Xia, P. Papadopoulos, S. N. Chakravarthula, B. Xiao, M. Van Segbroeck, J. Kim *et al.*, "Automated evaluation of non-native english pronunciation quality: combining knowledge-and data-driven features at multiple time scales." in *Proc. Interspeech*, 2015, pp. 493–497.

[22] S. Silva and A. Teixeira, "Unsupervised segmentation of the vocal tract from real-time mri sequences," *Computer Speech & Language*, vol. 33, no. 1, pp. 25–46, 2015.

[23] P. J. Jackson and V. D. Singampalli, "Statistical identification of articulation constraints in the production of speech," *Speech Communication*, vol. 51, no. 8, pp. 695 – 710, 2009.

[24] C. P. Browman and L. Goldstein, "Gestural specification using dynamically-defined articulatory structures," *Journal of Phonetics*, vol. 18, pp. 299–320, 1990.

[25] P. Boersma, "Praat: doing phonetics by computer [computer program]," *http://www. praat. org/*, 2020.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Analysis of Visual Features for Continuous Lipreading in Spanish

David Gimeno-Gómez , Carlos-D. Martínez-Hinarejos

Pattern Recognition and Human Language Technologies Research Center,
Universitat Politècnica de València, Camino de Vera, s/n, 46022, València, Spain

dagigo1@dsic.upv.es, cmartine@dsic.upv.es

Abstract

During a conversation, our brain is responsible for combining information obtained from multiple senses in order to improve our ability to understand the message we are perceiving. Different studies have shown the importance of presenting visual information in these situations. Nevertheless, lipreading is a complex task whose objective is to interpret speech when audio is not available. By dispensing with a sense as crucial as hearing, it will be necessary to be aware of the challenge that this lack presents. In this paper, we propose an analysis of different speech visual features with the intention of identifying which of them is the best approach to capture the nature of lip movements for natural Spanish and, in this way, dealing with the automatic visual speech recognition task. In order to estimate our system, we present an audiovisual corpus compiled from a subset of the RTVE database, which has been used in the Albayzín evaluations. We employ a traditional system based on Hidden Markov Models with Gaussian Mixture Models. Results show that, although the task is difficult, in restricted conditions we obtain recognition results which determine that using eigenlips in combination with deep features is the best visual approach.

Index Terms: lipreading, machine learning, speech technologies, computer vision, hidden markov models, deep learning

1. Introduction

During a conversation, our brain is responsible for combining information obtained from multiple senses in order to improve our ability to understand the message we are perceiving. Different studies have shown the importance of presenting visual information in these situations, as well as its relationship with the sounds produced. Principally, we stand out the studies carried out by McGurk and McDonald [1], where they demonstrated that if the mouth expression does not match with the emitted sound, the listener was confused, perceiving a sound different from what it really was. Nevertheless, lipreading is a complex task whose objective is to interpret speech when audio is not available. By dispensing with a sense as crucial as hearing, since this signal presents a greater amount of information regarding speech recognition, it will be necessary to be aware of the challenge that this lack presents.

Our ideal purpose is to build a system capable of imitating the human ability to interpret continuous speech by reading the lips of the speaker. Due to the absence of acoustic cues, some of the main challenges we have to deal with are visual ambiguities and silence modelling [2, 3]. Therefore, an essential factor is to identify a suitable representation that manages to capture the nature of lip movements and how it affects to the recognition quality. Consequently, the central core of this paper deals with an analysis of speech visual features [4, 5].

For this comparison, a fixed decoding system must be chosen. In our case, we employed a traditional approach to define the automatic system, in other words, a system based on Hidden Markov Models combined with Gaussian Mixture Models (GMM-HMM), an approach that has been widely used in Acoustic Speech Recognition (ASR) [6]. Although this is not the state-of-the-art for speech-related signal recognition, it is an appropriate option for comparing the different possibilities for feature extraction. Unlike in ASR, when we deal with Visual Speech Recognition (VSR) our basic speech unit is not the phoneme, but the one known as the viseme, which is associated with the representation of the phoneme on the visual domain [7]. Unfortunately, there is not direct or one-to-one correspondence between them, which causes visual ambiguities. An interesting work [2] describes a phoneme-to-viseme mapping for Spanish and concludes its usefulness compared to recognition through phonemes directly. However, we decided to establish the phoneme like basic speech unit in our work as many authors have done [8, 9, 10].

Apart from that, in order to estimate our system, an audiovisual corpus focused on continuous speech has been built from a subset of the RTVE database [11], with Spanish being the language to be interpreted. The RTVE database is a well-known database which has been used in the Albayzin evaluations [12]. Taking into account all these aspects, we can integrate our task both in the field of Speech Technologies and Computer Vision.

In relation to the rest of the paper and its organization, we mention that Section 2 provides the context and historical evolution around the task of VSR or Automatic Lipreading Recognition (ALR). Section 3 presents several details regarding the built audiovisual corpus. Then, Section 4 describes the different visual approaches considered in order to represent the nature of lip movements. Section 5 shows the experimental process carried out in our work, as well as certain insights and comments regarding our results. Finally, conclusions and future lines of research are offered in Section 6.

2. State of the art

In its origins, automatic speech recognition systems focused only on acoustic information, since this signal is more informative to distinguish phonemes [13]. Nowadays, these models are powerful systems capable of understanding spoken language with great quality [14]. However, when the acoustic signal is damaged or corrupted, the performance of these systems decline considerably [13, 10]. Therefore, many authors have studied how the incorporation of visual cues alongside acoustic information cause a significant improvement over interpretations supplied by the system in these situations [10, 15]. Additionally, several studies related to Silent Speech Interfaces (SSI) [16] were carried out to deal with the possible absence of the acoustic signal in the field of Speech Technologies.

In this way, in the last decades there has been an increase in the interest of decoding speech using exclusively the informa-

10.21437/IberSPEECH.2021-47

tion from the visual channel. As Fernandez-Lopez and Sukno suggest in their review [13], advances achieved by this type of systems have been conditioned, among other reasons, by the evolution reflected over available audiovisual databases. These databases began by tackling simple tasks from alphabet and digit recognition, such as AVLetters [4] and CUAVE [17] corpora. More recent datasets provide the necessary support to estimate approaches in charge of interpreting spontaneous speech, for instance, *Lipreading Sentences in the Wild* (LRS) [18]. Regarding Spanish, we stand out the VLRF corpus [8], despite the fact that it differs from our objectives by recording the scenes under controlled conditions and ensuring that speakers strain theirselves to vocalize adequately and expressively. Lately, the CMU-MOSEAS database [19] has been compiled, among other languages, for Spanish. This is an interesting corpus, as it provides a multimodal point of view, supplying information related with the emotions and subjectivity expressed by the speaker.

At the beginning, these tasks were developed under a traditional paradigm, that is, mainly through the well-known HMMs. Since this is our case, we highlight some publications, such as studies carried out by Thangthai, Cox, and Howell, among others [20, 9], where they employed an HMM per phoneme, evaluating both dependent and context-independent models. On the other hand, in relation to Spanish, we mention again the paper developed by Fernandez-Lopez and Sukno [8], where we emphasize their study regarding recognition at the phoneme level over the VLRF corpus. However, the research has gravitated towards Deep Learning technologies. More concretely, end-to-end architectures, formed by combining Long Short Term Memory (LSTM) [21] and Convolutional Neural Networks (CNN) [22], have been the most widely used topology. In fact, Zisserman [18] reached the state-of-the-art in continuous VSR by employing this approach and achieving an error rate of around 50% at word level.

3. Audiovisual corpus

As we mentioned above, we have compiled an audiovisual corpus focused on the task of continuous lipreading recognition, where we could find a large number of speakers in a wide range of scenarios, including variations on intra-personal aspects or light conditions. We compiled it from a subset of the RTVE database [11] which has been employed in Albayzín evaluations [12]. The RTVE database is made up of different programs broadcasted by *Radio Televisión Española* but, in our work, we compiled the corpus only from the news program 20H broadcast by the *Canal 24 horas*. In this program, we have selected scenes where a unique speaker talks from different distances to the camera and in diverse scenarios, either inside a record studio or in outdoor locations. Furthermore, the speaker does not always maintain a frontal plane but can sometimes adopt tilted postures. Other details regarding the compiled dataset are shown in Table 1.

Table 1: *Details regarding the compiled audiovisual corpus.*

Language	Spanish		
Resolution	480×270 pixels, 30 frames/second		
Speakers	57	**Males:** 17	**Females:** 40
Duration	∼3 hours		
Utterances	2792		
Vocabulary	2885		
Running words	∼35k		
Phonemes	23		
Words per utterance	**Median:** 10	**Max:** 62	**Min:** 1
Phonemes per utterance	**Median:** 46	**Max:** 270	**Min:** 5

Before continuing, it is necessary to mention that the resolution of 480×270 pixels refers to the full record scene. Our region of interest, that is, the speaker's mouth, is of variable dimensions. Therefore, we stablished its size in 32×16 pixels.

4. Speech visual features

In the literature, a large number of approaches regarding the visual representation of speech have been studied. Many authors [4, 23, 24] have employed traditionals techniques to extract these features, such as Principal Component Analysis (PCA), Active Appearence Models (AAM), or Optical Flow. Moreover, other authors have delegated the responsibility of extracting visual features on neural networks [25, 26, 5], and more specifically on Convolutional Autoencoders.

However, there is no consensus or agreement in the literature on what is the best option for the extraction of visual speech features. Therefore, in our work we studied three types of features that we describe in subsections from 4.1 to 4.3. In all the approaches, the use of the OpenCV library [27] and the Dlib toolkit [28] allowed us to identify 68 facial landmarks. From some of these landmarks, as left part of Figure 1 reflects, we were able to extract our region of interest.

4.1. Geometric features

In this first approach, the study of continuous sign language carried out by Hermann Ney and other authors [29] is taken as a reference. In this way, we defined, thanks to the location of landmarks described above, a set of 18 high-level features, such as width, height, or area of speaker's mouth. However, when the speaker is more or less close to the camera the same metric (for example, mouth's width) would acquire a value in different magnitudes, even in the case of the same mouth posture or physiognomy. For this reason, we decided to locate a more stable region where each of the measured distances can be properly normalized. This is the region highlighted by a larger blue rectangle on the right side of Figure 1. Thus, the resulting geometric features are normalized using the size of this area.

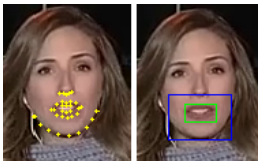


Figure 1: *Aspects regarding geometric features.*

4.2. Eigenlips

This concept is influenced by the studies carried out on facial recognition [30]. Then, as other authors did [3], after computing PCA over our training set we could obtain the eigenlips shown in Figure 2. The first component, as suggests the image, stands out the lip corners, since these are the parts that suffer the greatest deformation throughout a speech. As for the rest of eigenlips, some emphasize lip contours while others highlight zones where we can find teeth or tongue. These are aspects that we can not reach when we work with pure geometric features, but that are of vital importance for visual speech recognition.

Figure 2: *Eigenlips obtained after applying PCA.*

Another important issue is that, with the intention of making easier the extraction of these and the features described in Subsection 4.3, we apply a mouth alignment. In other words, as Figure 3 suggests, we rotate those regions of interest where we observe that the speaker's mouth is tilted or inclined.

Figure 3: *Schematic process of mouth alignment.*

4.3. Deep features

The last approach, as we mentioned above, is based on Convolutional Autoencoders [31], a neural network whose main purpose consists of reconstructing the image received as input from an abstract and compact representation which has been obtained previously from the original image. Once this statistical model is trained, we can dispense with decoder because it is the encoder the component we need to extract our visual features. This scheme is reflected in Figure 4.

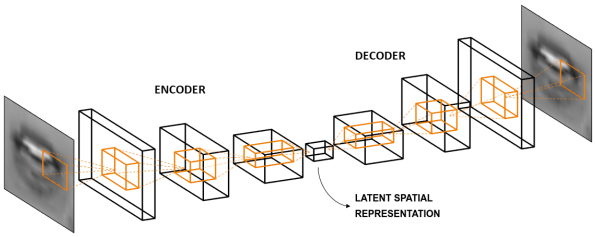


Figure 4: *Scheme of the proposed Convolutional Autoencoder.*

The encoder architecture is entirely based on that presented in [25], since they employ low-resolution images as is our case. Thus, we were able to obtain high quality reconstruction results, as shown in Figure 5. Finally, we remind that in these features we apply a mouth alignment too.

Figure 5: *Reconstructions examples obtained by the defined Convolutional Autoencoder. For each images block, left column: original image; right column: reconstruction.*

5. Experiments

All our experiments were performed with the Kaldi toolkit [32]. This toolkit provides the support to build high performance systems focused on Speech Technologies, from traditional approaches to hybrid models or systems based entirely on Deep Learning. In our case, as we stated in the introduction, we employed a traditional GMM-HMM system. This system is formed by three modules: the Optical Model, in charge of interpreting the pronounced phonemes from visual features sequence; the Lexical Model, responsible for building the words from the phonemes provided by the previous module; and the Language Model, capable of combining the words provided by the Lexical Model with the intention of generating the message interpreted by the system. On the other hand, our corpus is divided into two partitions, allocating to the test set those speakers who did not reach a minimum of seconds. Thus, we get a train set composed of 2672 utterances emitted by 43 different speakers, reaching around 3 hours of data, whereas the test partition comprises 120 samples from 13 speakers, covering 0.13 hours of utterances. Then, due to the limited amount of available data, we had to estimate a context-independent system, also know as monophonic system.

At the beginning, we carried out several speaker-independent experiments with an Open Language Model, but because of the difficulty of the task, we did not achieve minimally acceptable results (error rates greater than 90%). Consequently, it was necessary to make experiments in a more constrained scenario in order to obtain conclusions on the use of the different features. Therefore, we decided to relax the task complexity by employing a Closed Language Model. In other words, a Language Model estimated only from the text included in the test partition. In this way, the system has a reduced set of alternatives when it interprets the message, which allows us to focus our experiments on the performance of the different types of features. In fact, as suggested in [8], an acceptable recognition at phonemic level does not necessarily imply a good quality performance when word level decoding message is carried out.

Our first experiment focused on studying HMM's topology, one of the most relevant factors in relation to temporal alignment of visual data. The classic topology in ASR (3 states left-to-right with self-loops) provided us a poor recognition rate. Then, employing raw geometric features, we tested several topologies. After these experiments, we observed that if we reduce the number of states and we add transitions of each of these states to the final state, system performance obtains, in general, a considerable improvement. According to these results, we believe that this behaviour may be caused by the limited frequency of information (30 frames/second) that presents the visual channel with respect to the standard representation (Mel Frequency Cepstral Coefficients, MFCC) of acoustic data (100 frames/second) [6]. Consequently, in the rest of our work we employed the topology shown in Figure 6: 2 states left-to-right with self-loops and skip transitions.

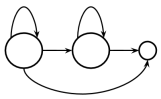


Figure 6: *HMM's topology employed in our experiments.*

Once this aspect is specified, we can address the analysis of the visual speech features with the intention of determining which of them, either isolated or combined with each other, is the best representation to capture the nature of lip movements. In addition, we study how the incorporation of temporal or delta-delta coefficients, used in ASR [6], affects to the recognition quality. More precisely, we study different contexts, both a greater and a lower scope. On the other hand, as Lan presented in [23], we have observed how applying a z-score normalization causes a considerable recognition quality improvement. Then, in our analysis we studied two types of normalization on all the features presented in Section 4:

- Normalization per speaker: all the utterances of a speaker are taken in the normalization; this is aimed at mitigating the differences that may exist in the aspect of the speaker in his/her different utterances (light conditions, facial hair, lipstick colour, temporal scars or marks in the mouth, ...).
- Normalization per utterance: the normalization is for each single utterance; this is done with the intention of

Table 2: *Results (WER) for visual speech features with speaker and utterance normalization.* ***raw****: refers to the raw features, without adding any type of temporal coefficient.* $\Delta\Delta_x$*: applies on the features the coefficients* delta-delta *with a context of x frames. Best result for each set of features and normalization in bold.*

	***z-score* normalization per speaker**				***z-score* normalization per utterance**			
Features	**raw**	$\Delta\Delta_1$	$\Delta\Delta_2$	$\Delta\Delta_3$	**raw**	$\Delta\Delta_1$	$\Delta\Delta_2$	$\Delta\Delta_3$
geometricFeats	51.3±8.5	49.2±8.2	**37.7±8.2**	50.5±7.8	46.1±8.8	49.8±8.6	**36.8±7.1**	48.7±8.0
eigenLips	71.6±8.6	60.6±8.5	**57.1±8.3**	65.8±7.5	66.6±8.8	**49.9±8.1**	55.6±8.5	53.0±8.4
deepFeats	46.6±8.5	**31.7±7.3**	35.9±7.7	39.1±7.8	72.4±7.9	58.9±8.2	54.8±8.9	**54.7±8.2**
geometricFeats+eigenLips	29.6±7.5	30.3±6.5	34.2±7.2	**26.6±6.0**	**26.1±6.2**	34.6±7.1	31.2±6.6	34.9±6.4
geometricFeats+deepFeats	32.9±7.9	**29.8±7.1**	34.1±7.0	41.3±6.7	**29.3±7.7**	36.5±7.0	33.0±6.9	41.3±7.6
eigenLips+deepFeats	45.2±8.2	33.6±7.9	**23.7±6.4**	35.4±7.1	38.0±7.5	29.6±6.8	**26.8±7.2**	31.0±7.2
geometricFeats+eigenLips+deepFeats	30.4±6.8	**27.3±6.5**	33.4±6.5	41.5±6.7	**34.6±7.7**	36.4±6.7	**34.6±6.8**	43.0±6.8

reducing the differences among the different utterances from different speakers (i.e., to obtain speaker independency) and conditions (e.g., focus variability).

All the results are evaluated by the well-known Word Error Rate (WER) with 95% confidence intervals obtained by the boostrap method described in [33].

Results are presented in Table 2. First, we stand out that the incorporation of delta-delta coefficients cause, in general, an improvement on system performance. Of course, depending on the type of features or normalization, it is convenient to use one context or other.

On the other hand, if we focus only on those experiments that study the visual features individually, we conclude that deep features are the best representation, in isolated way, to address VSR, as long as we normalize per speaker. In contrast, when we apply a normalization per utterance, we notice how the quality of these features suffer a drastic deterioration, while the rest of visual approaches improve their results. This may mean that deep features, as they are highly dependent on the pixel values, are more affected by changes in light conditions or certain intra-personal aspects. In this way, if these features are processed by a normalization per speaker, they are more benefited in order to interpret speech visually. In contrast, geometric features are more dependant on the location of the specific pixels, and thus utterance normalization fits better for these features. Eigenlips depend on both specific pixels and their values, which makes it to be the worst option when normalizing by speaker and to not improve geometric features when normalizing by utterance.

Regarding the experiments that explore the combinations of visual features, we confirm that, as a general rule, feature combination produces a decrease in error rates. Therefore, we can deduce that the studied features complement each other and manage to provide a more robust representation. Nevertheless, this is not always the case; in certain occasions these results are overlapped with the best error rates obtained in experiments where features were employed individually. On the other hand, the eigenlips and deep features combination, if we normalize per speaker, is established as the best approach to address the automatic lipreading, reaching around 23.7% WER, although differences are not significant with respect to only using deep features. Seemingly, thanks to appearance aspects contained in eigenlips and the great potential that deep features demonstrated regarding mouth physiognomy reconstruction, a high quality representation has been achieved. However, it is worth noting that if we do not incorporate delta-delta coefficients, we can verify how the geometric features and eigenlips combination improves the performance of the previously mentioned approach.

Finally, we stand out that the combination of all the features analyzed in our work forms a representation with a large dimension. This fact can cause difficulties when modelling data statistically. Consequently, unlike most cases in which isolated features are used, introducing temporal coefficients does not always imply better results.

6. Conclusions

In this paper, an extensive study has been carried out regarding visual speech features in order to address an automatic lipreading task for natural Spanish. Therefore, in addition to compiling a preliminary audiovisual corpus, approaches based on both traditional techniques and Deep Learning architectures have been addressed to represent the nature of lip movements. After our experiments, we conclude that the combination of eigenlips and deep features, as long as we apply certain aspects such as normalization and temporal coefficients, provide the best approach to interpret speech visually. On the other hand, aspects in relation to temporal alignment of visual data have been studied. More concretely, according to the results obtained in this work, the HMM's topology has been modified regarding standards in acoustic speech recognition.

However, visual speech recognition remains an open problem. Lipreading is a complex task where researchers have to face with visual ambiguities and other issues such as silence modelling. In addition, there is not a consensus or agreement regarding a suitable visual speech representation. For these and others reasons, further research is necessary. Authors such as Fernandez-Lopez and Sukno [13] have suggested that future lines of research should be developed around temporal alignments and context modelling. On the other hand, we consider to shift our study towards a pure Deep Learning approach. More precisely, we aim at an end-to-end architecture whose parameters, including those in charge of extracting visual features, are estimated according to the mistakes identified in the message decoding stage. In other words, we believe that this direct learning could be more useful than addressing the task with a traditional approach, where each module is independent of the other. In order to achieve this objective, we must increase the number of seconds which forms the compiled audiovisual corpus. In this way, we expect to get a large amount of data which represents the nature of natural speech and be able to estimate our statistical models appropriately. Finally, it would be interesting to study whether a suitable viseme-phoneme correspondence for Spanish can lead to advances in the matter.

7. Acknowledgements

This work was partially supported by Generalitat Valenciana under project DeepPattern (PROMETEO/2019/121) and by Ministerio de Ciencia under project MIRANDA-DocTIUM (RTI2018-095645-B-C22).

8. References

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[2] A. Fernandez-Lopez and F. M. Sukno, "Optimizing phoneme-to-viseme mapping for continuous lip-reading in spanish," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics*. Springer, 2017, pp. 305–328.

[3] K. Thangthai, "Computer lipreading via hybrid deep neural network hidden markov models," Ph.D. dissertation, University of East Anglia, 2018.

[4] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, 2002.

[5] D. Parekh, A. Gupta, S. Chhatpar, A. Yash, and M. Kulkarni, "Lip reading using convolutional auto encoders as feature extractor," in *5th International Conference for Convergence in Technology (I2CT)*. IEEE, 2019, pp. 1–6.

[6] M. Gales and S. Young, *The application of hidden Markov models in speech recognition*. Now Publishers Inc, 2008.

[7] C. G. Fisher, "Confusions among visually perceived consonants," *Journal of speech and hearing research*, vol. 11, no. 4, pp. 796–804, 1968.

[8] A. Fernandez-Lopez, O. Martinez, and F. M. Sukno, "Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database," in *12th International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 208–215.

[9] K. Thangthai, R. W. Harvey, S. J. Cox, and B.-J. Theobald, "Improving lip-reading performance for robust audiovisual speech recognition using dnns." in *AVSP*, 2015, pp. 127–131.

[10] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[11] E. Lleida, A. Ortega, A. Miguel, V. Bazán, C. Pérez, M. Zotano, and A. de Prada, "Rtve2018 database description," *Vivolab and Corporación Radiotelevisión Española, Zaragoza, Spain*, 2018, [Online] Available: http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf.

[12] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. de Prada, "Albayzin 2018 evaluation: the iberspeech-rtve challenge on speech technologies for spanish broadcast media," *Applied Sciences*, vol. 9, no. 24, p. 5412, 2019.

[13] A. Fernandez-Lopez and F. M. Sukno, "Survey on automatic lip-reading in the era of deep learning," *Image and Vision Computing*, vol. 78, pp. 53–72, 2018.

[14] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[15] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE transactions on multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[16] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[17] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Cuave: A new audio-visual database for multimodal human-computer interface research," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 2002, pp. II–2017–II–2020.

[18] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3444–3453.

[19] A. B. Zadeh, Y. Cao, S. Hessner, P. P. Liang, S. Poria, and L.-P. Morency, "Moseas: A multimodal language dataset for spanish, portuguese, german and french," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 1801–1812.

[20] D. Howell, S. Cox, and B. Theobald, "Visual units and confusion modelling for automatic lip-reading," *Image and Vision Computing*, vol. 51, pp. 1–12, 2016.

[21] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with lstm," in *Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, vol. 2, 1999, pp. 850–855 vol.2.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[23] Y. Lan, B.-J. Theobald, R. Harvey, E.-J. Ong, and R. Bowden, "Improving visual features for lip-reading," in *Auditory-Visual Speech Processing*, 2010, pp. 142–147.

[24] A. A. Shaikh, D. K. Kumar, W. C. Yau, M. C. Azemin, and J. Gubbi, "Lip reading using optical flow and support vector machines," in *3Rd international congress on image and signal processing*, vol. 1. IEEE, 2010, pp. 327–330.

[25] I. Fung and B. Mak, "End-to-end low-resource lip-reading with maxout cnn and lstm," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2511–2515.

[26] K. Paleček, "Extraction of features for lip-reading using autoencoders," in *International Conference on Speech and Computer*. Springer, 2014, pp. 209–216.

[27] G. Bradski, "The opencv library," *Dr Dobb's J. Software Tools*, vol. 25, pp. 120–125, 2000.

[28] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[29] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, pp. 108–125, 2015.

[30] K. Delac, M. Grgic, and P. Liatsis, "Appearance-based statistical methods for face recognition," in *Proceedings of the 47th International Symposium ELMAR focused on Multimedia Systems and Applications, Zadar, Croatia*, 2005, pp. 151–158.

[31] Y. Bengio, "Learning deep architectures for ai," *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[32] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[33] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in asr performance evaluation," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. 409–412.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Implementation of neural network based synthesizers for Spanish and Basque

Víctor García Romillo, Inma Hernáez Rioja, Eva Navas

HiTZ Center - Aholab, University of the Basque Country (UPV/EHU)
victor.garcia@ehu.eus, inma.hernaez@ehu.eus, eva.navas@ehu.eus

Abstract

This paper describes the implementation of neural-network based Text-to-Speech (TTS) synthesizers for Spanish and Basque. In order to develop this research, the voices of one male and one female speakers, both bilinguals, are used in a data set of around 4 and a half hours for each voice and language. The system uses Tacotron to compute mel-spectrograms from the input text sequence and Waveglow to obtain the resulting audios.

Training the mentioned models with a limited amount of data leads to synthesis errors in some utterances, affecting the naturalness of the audios and even producing unintelligible speech. In this paper, we describe the method followed to automatically detect erroneously synthesized audios and the strategy followed to address the causes of the errors. The designed method has been validated by testing the TTSs using a large set of out-of-domain sentences. In the end a fully operational system is developed, with capacity to generate good quality and natural audios, as showcased by the evaluation conducted.

Index Terms: speech synthesis, robustness, text to speech, Basque, Spanish

1. Introduction

Text-to-Speech (TTS) systems transform input written language into synthetic speech. Traditionally, two main approaches have been used: unit selection (US) based concatenative synthesis and statistical parametric (SP) speech synthesis. The former uses big databases segmented into sub-word units, and attempts to select the best sequence of them to match the target sentence [1, 2]. The latter approach generates mathematical models that attempt to relate input text features with acoustic features [3, 4]. Hybrid approaches have been also proposed, with the intention of combining the segmental naturalness of US and the consistency offered by SP [5, 6].

Nowadays, deep neural network (DNN) based systems are state-of-the-art in speech synthesis [7, 8]. Neural networks have benefited the speech synthesis field by improving the quality and naturalness of the synthetic voices with respect to the traditional systems. Another contribution made by neural networks is the possibility of training and designing the systems in an end-to-end (E2E) fashion. While traditional multi-stage pipelines are complex and require from large domain expertise, E2E systems reduce the complexity by extracting the audio directly from the input text without needing separated models.

There are different neural network based E2E architectures for TTS [9, 10]. The TTS systems built in this work are based on Tacotron 2 [11]. Tacotron 2 is a sequence-to-sequence [12] architecture that maps character embeddings to mel-scale spectrograms. To transform the output spectrograms into waveforms, we use WaveGlow [13]. WaveGlow is a neural vocoder that combines insights of WaveNet [14] and Glow [15] to produce high-quality audio using parallel capabilities of GPUs.

Although end-to-end TTS systems have shown excellent results in terms of audio quality and naturalness, there are some issues to be faced. On the one hand, these systems usually suffer from low training efficiency, requiring a sizable set of text and audio pairs to train properly. On the other hand, synthesized speech is usually not robust, due to alignment failures between input text and speech during the generation.

In this paper we describe the implementation of four TTS systems, two for Spanish and two for Basque based on the previously mentioned architectures. To evaluate them, several out-of-context utterances were synthesized. We propose a method to automatically detect sentences where a poor alignment leads to unintelligible synthetic speech. We also describe the strategy followed to address the causes of the issues found during the evaluation of the initial implementation. Finally, we conduct an evaluation to check if the changes improve the robustness of the models while maintaining good quality and naturalness over the synthetic voices.

The paper is organized as follows. Section 2 describes the data used to train and evaluate all models. Section 3 contains a description of the construction of the systems, along with an analysis of their issues and the approaches taken to address them. Subjective and objective evaluations of the final implementation are shown in section 4. Finally, some conclusions are drawn.

2. Materials

In order to train the models, datasets containing speech signals and their corresponding transcriptions are needed. For evaluation purposes we only required the text. The following sections describe all the data used in this work and the processing applied to it.

2.1. Training dataset

The neural-network approach taken in the construction of the TTS system demands having available a speech corpus with its corresponding transcriptions. In this work we have used two phonetically balanced corpora of about 4000 sentences, one for Basque and one for Spanish, which have been recorded by one male and one female bilingual speakers. Figure 1 shows a distribution of the amount of words per utterance in the Spanish and Basque corpora. The datasets have an average of 12.84 ± 3.93 and 10.10 ± 2.81 words per sentence in Spanish and Basque. Each recorded corpus has a duration of approximately 4 hours and 30 minutes.

In order to obtain faster convergence times along with higher synthesis quality, speech signals and their corresponding transcriptions needed to be processed. Regarding the audio, silences at the beginning of the sentences were removed and silences at the end were trimmed to 150ms, as this process eases the learning of the alignment between text and audio. The original sampling frequency of 48000 Hz was decimated to

10.21437/IberSPEECH.2021-48

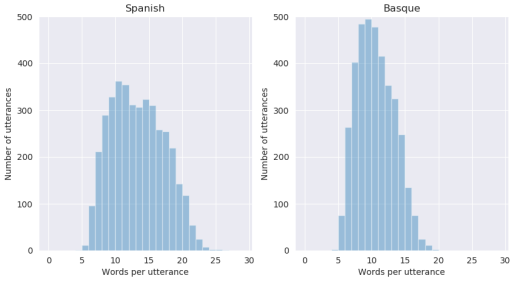


Figure 1: *Distribution of words per utterance in the Spanish and Basque corpora used for training*

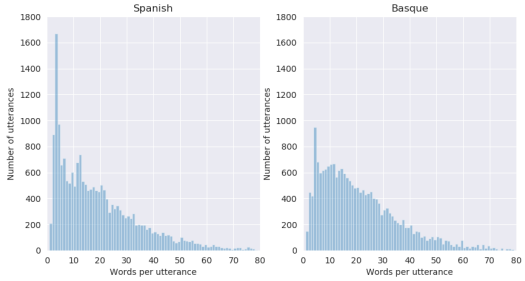


Figure 2: *Distribution of words per utterance in the Spanish and Basque corpus used for robustness evaluation*

22050 Hz to match the sample rate used in a pre-trained model that would be later adapted with the Basque and Spanish voices. Finally, recordings longer than 9 seconds were removed to avoid convergence issues due to memory restrictions from the available GPUs.

In relation to the transcriptions, the first step was standardizing the encoding of all texts to UTF-8. A linguistic Front-End in Spanish and Basque [16] was used to normalize and clean the utterances. The same Front-End was used to extract the phonetic transcriptions of all the utterances expressed in SAMPA alphabet [17].

2.2. Testing datasets

For the evaluation of the deployed TTS systems two different written text datasets were used. To evaluate the robustness of the systems a dataset containing out-of-domain sentences was used. The evaluation of the quality and naturalness of the synthetic voices was conducted using in-domain sentences.

a) Parliamentary texts: This corpus was provided by the MintzAI project [18] [1]. The utterances in the corpus were obtained from transcriptions of the Basque Parliament sessions, both in Basque and Spanish. Being transcriptions of spoken parliamentary speeches, the complexity of the sentences differ greatly from that of the sentences in the training dataset. Also, the utterances have varying lengths from a few words to very long sentences, a particularly difficult scenario for the Tacotron 2 model [19]. From the complete dataset we randomly selected 20000 sentences. Figure 2 shows the distribution of words per utterance in this dataset. As it can be seen, the distribution in this dataset differs from the one used in training. Testing datasets have an average of 19.46 ± 16.81 and 21.27 ± 16.55 words per sentence in Spanish and Basque.

b) Texts from novels and tales: This corpus contains sentences extracted from different tales and novels. It is a phonetically balanced corpus with 450 sentences in Spanish and Basque. As this corpus is used to evaluate the quality and naturalness of the synthetic signals, we used sentences with a length distribution similar to the one shown in Figure 1. The motivation behind this choice was preventing synthesis failures due to alignment issues in long sentences.

The processing applied to both corpora was the same as the one applied to the transcriptions of the training corpora.

[1] http://www.mintzai.eus/indice.html

3. Methodology

In this section we describe the procedure followed in the development of the final neural TTS systems. As will be described in subsection 3.1, we started with an initial implementation of the systems using slightly modified versions of the reference architectures described in the literature. However, the system suffered from some issues that will be described in subsection 3.2. Subsection 3.3 describes the steps taken towards a final implementation that aims to solve the previously mentioned issues.

3.1. Initial implementation

The TTS architecture used in this work consists of two components: a feature prediction network that predicts mel-spectrogam frames from the input text, and a neural vocoder able to generate speech from mel-spectograms. The initial implementation of the developed TTS system uses a Tacotron 2 [11] based model as feature prediction network and Waveglow [13] as neural vocoder. Other neural vocoders like Wavenet [14] and MelGAN [20] were tested, but Waveglow was chosen as it offers a high quality voice with easy training and fast synthesis times.

A restrictive issue when it comes to training Tacotron is the high data volume demand. According to [21], the best audio quality is obtained when using between 10 and 40 hours of data, whereas using less data still produces good quality albeit with certain degradation. As we do not have such big amount of high quality transcribed data available, we opted to apply transfer learning over a publicly available pre-trained model provided by NVIDIA [22]. This model was trained with LJSpeech dataset [23], an English corpus with approximately 24 hours from a single female speaker. As the main objective of this work was developing a TTS system with male and female voices in Spanish and Basque, 4 different models were trained using the training database described in section 2.1. The training of the 4 models converged after 15k steps, lasting approximately one day for each model with a NVIDIA TITAN RTX graphics card and batch size of 64. To prevent over-fitting, attention layer dropout was set to 0.4 and decoder dropout rate was set to 0.1. Learning rate remained constant through the whole training at 0.001.

Regarding the neural vocoder, a pre-trained model was also used due to the computational cost of training a new model from scratch. The obtained models produced good quality voices, but a few issues were identified when synthesizing speech with them. The following section covers all of them.

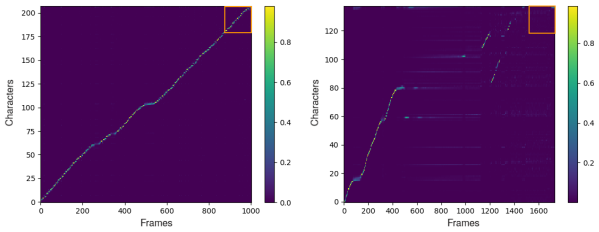


Figure 3: *Correctly aligned (left) and badly aligned (right) sentence alignments*

3.2. Issues found on the initial implementation

To evaluate the trained models, a set of unseen sentences were synthesized. These sentences are obtained from dataset a) described in section 2.2. The synthesized signals displayed two main issues: poorer naturalness for male voices and lack of robustness.

3.2.1. Poorer naturalness for male voices

A first informal listening test of the synthesized speech signals allowed to detect that, regardless of the input text, the naturalness for the two male voices (Basque and Spanish) was poorer than that of the female voices, showing unpleasant noises and buzziness. Visual inspection of the mel-spectrograms generated for female and male voices showcased no major degradation in the latter. We deemed this error to the fact that the Waveglow's pre-trained model corresponds to a female voice. This effect has also been reported in [24].

3.2.2. Lack of robustness

Closer inspection over the output attention graphs of the synthesized utterances displayed that, occasionally, the model was failing to identify the generation stop token. This results in noisy fragments at the end of the generated audio signals. In addition to this, some attention graphs also showed that the model was failing to attend to the correct input character at certain decoding steps, resulting in speech generation instability. As stated in [25, 19], autoregressive attention-based systems are prone to alignment instability, causing word skippings, repetitions and in the worst cases unintelligible speech.

In order to measure the number of audio files affected by either the noise at the end or the lost of alignment during synthesis, a post processing stage was added to the model. The function of this stage is to check the last decoding steps of the inference process, verifying that the attention matrix weights given to the last characters are higher than a threshold. Figure 3 shows the alignment matrix of 2 different synthesized sentences, with the input characters at the vertical axis and the output frames at the horizontal axis. The left image shows a correct alignment and the right image shows that the attention has been lost midway. The figure shows in a rectangle the area that corresponds to the final decoding steps. The algorithm checks row-wise the assigned weights in the selected area, searching for values over a threshold. The values for the area size (10x50) and the threshold (0.3) for the weights have been chosen empirically.

3.3. Final implementation

Addressing the aforementioned issues was crucial to develop natural voices in Basque and Spanish in a robust manner. The quality difference between the female and male voices was large and the attention instability during synthesis made the model unreliable for the task of generating a large amount of synthetic sentences without supervision.

The amount of available data did not suffice to apply transfer learning over the pre-trained neural vocoder model. Nonetheless, as stated in [26], Waveglow is able to generalize to unseen speakers. This feature enables the use of additional corpora containing male voices to improve the quality of the synthesized waveforms. In this case, a single speaker male voice extracted from an audiobook in Spanish was used. The final length in Spanish and Basque resulted in approximately 19 hours of male voices. The available memory in the GPU used for training forced to set the batch size to 3. The learning rate remained constant at a value of 0.001 and the training converged after 18 days using a NVIDIA TITAN RTX graphics card.

Regarding the lack of robustness observed in Tacotron 2 different approaches can be applied. Some studies propose modifying the existing attention mechanism aiming to improve the model instability, as related in [27, 28]. Other technique is based on injecting prior knowledge into the model to improve the training of the existing attention mechanism. We opted for the latter, specifically by implementing pre-alignment guided attention [25] as it improves the model stability when synthesizing long utterances, while also enhancing the training efficiency.

The basic idea of pre-alignment guided attention is introducing an explicit target to guide the model to learn the attention during the training process. These targets are time-aligned phoneme sequences. In order to obtain them, the linguistic Front-End for Spanish and Basque was used along with a forced alignment tool (Montreal Forced Aligner [29]).

Once all the necessary input files were obtained and the model was modified to include the new attention loss metric, the training of the 4 models was done using the same hyperparameters as in the initial implementation. All the models converged after 12k iterations and the training lasted for approximately 12 hours.

4. Evaluation

In this section the robustness of the implemented models is evaluated, along with a Mean Opinion Score (MOS) evaluation of the quality and naturalness of the final implementation. Furthermore, we also evaluate the naturalness of the synthetic signals through a deep learning based assessment proposed by [30] called "Non-Intrusive Speech Quality Assessment" (NISQA) for TTS.

4.1. Robustness

To evaluate the robustness of the models we conducted a test where 20000 utterances were synthesized. These utterances are obtained from corpus a) described in section 2.2. The error detection algorithm proposed in section 3.2.2 was used to detect the files with critical synthesis errors. The relative improvement from the initial to the final implementation in terms of number of generation errors is shown in Table 1. As it can be seen, there is an important improvement in all cases.

4.2. MOS

The quality and naturalness of the final implementation of the systems were evaluated in a Mean Opinion Score test where participants had to rate both aspects in a 5-point scale. Utter-

Table 1: *Number of sentences with errors and relative improvement in percentage*

	Initial	Final	Improvement
Female Spanish	1791	1103	38.41
Female Basque	2596	1941	25.23
Male Spanish	1077	95	91.18
Male Basque	1206	274	71.31

ances from the dataset described in section 2.2 b) were used to generate the synthetic signals for the evaluation. As this dataset contains sentences in Spanish and Basque, bilingual subjects were required for the evaluation. Three different methods were used:

- Natural speech signals
- Synthetic speech signals generated with the DNN based TTS systems described in section 3.3
- Synthetic speech signals generated with the HTS based TTS systems previously developed in our research group [16][2]

Out of the 450 available sentences, each participant in the evaluation rated 6 randomly selected sentences per speaker (2), language (2) and method (3) (i.e. a total of 72 sentences). Overall, 33 subjects participated in the evaluation (only one among them was expert in speech technologies)[3].

Figure 4 shows the quality scores averaged for all models, languages and speakers together with 95% confidence intervals. In all cases the subjects conferred higher scores to the signals obtained with the Tacotron based systems than to those generated with the HTS systems. However the score is still lower than that of natural speech. On the other hand, female speech was rated higher than male speech in both languages for the DNN based systems. We deem this occurs because the neural vocoder training does not suffice to produce signals with the same quality for female and male voices. Furthermore, this preference is also shown for natural speech signals.

Figure 5 shows the averaged naturalness ratings of all models with 95% confidence intervals. As occurred in the quality evaluation, DNN based systems were also rated below natural speech but they scored higher than the HTS systems. Subjects showed preference for female voices over male voices in natural speech signals, and this also happened in the DNN based systems in both languages.

4.3. NISQA

NISQA-TTS [30] model is a speech naturalness estimator based on deep learning. According to the authors, it works language independently. Table 2 shows the scores provided by the NISQA-TTS model. As it can be seen, HTS based systems received more generous scores than the ones obtained in the MOS evaluation. Regarding the DNN based systems, NISQA-TTS model produces mixed results, being those more conservative in the case of female voices and more generous for male voices. In all cases DNN based systems score higher than HTS based ones, as happened in the MOS evaluation.

[2]https://sourceforge.net/projects/ahotts/

[3]Some examples can be found in http://aholab.ehu.eus/users/victor/IB2020.html

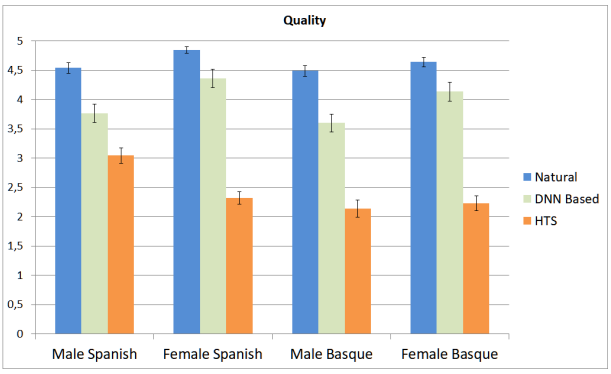


Figure 4: *Results of the quality assessment*

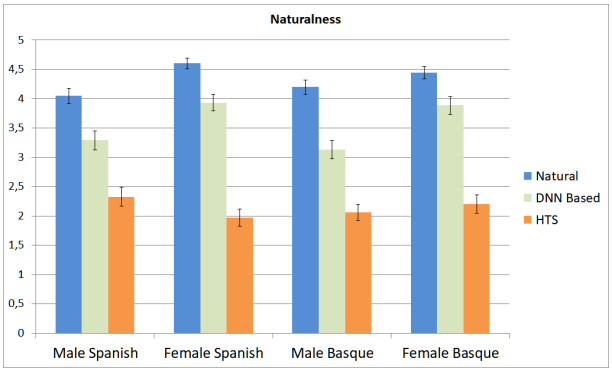


Figure 5: *Results of naturalness assessment*

Table 2: *NISQA evaluation scores with 95% confidence interval*

	HTS based	DNN based
Female Spanish	2.97 ± 0.05	3.34 ± 0.05
Female Basque	3.07 ± 0.04	3.46 ± 0.07
Male Spanish	3.62 ± 0.04	3.64 ± 0.08
Male Basque	2.95 ± 0.04	3.56 ± 0.08

5. Conclusions

This paper describes the implementation of several DNN based TTS systems for Spanish and Basque. Results from the conducted subjective and objective evaluation demonstrate that the robustness of the final systems improved and they are able to synthesize good quality and natural signals in both languages. The system is currently being used to generate training material to conduct speech to speech translation [18].

Future work includes improving the naturalness and robustness of the systems by increasing the amount of data used during the training. We also consider researching on different architectures for the feature prediction network to address the alignment issues without loosing quality.

6. Acknowledgements

This work has been funded by the Basque Government (Project refs. PIBA 2018-035, IT-1355-19 and MintzAI project KK-2019/00065).

7. References

[1] A. W. Black and P. Taylor, "Automatically Clustering Similar Units for Unit Selection Speech Synthesis," in *Proceedings of EUROSPEECH*. ISCA, 1997, pp. 601–604.

[2] N. Campbell and A. W. Black, "Prosody and the Selection of Source Units for Concatenative Synthesis," in *Progress in Speech Synthesis*. Springer New York, 1997, pp. 279–292.

[3] Y. J. Wu and R. H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proceedings of ICASSP*, vol. 1. IEEE, 2006.

[4] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, nov 2009.

[5] S. Tiomkin, D. Malah, S. Shechtman, and Z. Kons, "A hybrid text-to-speech system that combines concatenative and statistical synthesis units," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 5, pp. 1278–1288, 2010.

[6] I. Sainz, D. Erro, E. Navas, and I. Hernáez, "A hybrid tts approach for prosody and acoustic modules," in *Proceedings of INTERSPEECH*. ISCA, 2011, pp. 333–336.

[7] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proceedings of ICASSP*. IEEE, 2013, pp. 7962–7966.

[8] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "Tts synthesis with bidirectional lstm based recurrent neural networks," in *Proceedings of INTERSPEECH*, 2014, pp. 1964–1968.

[9] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of INTERPSEECH*. ISCA, 2017, pp. 4006–4010.

[10] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. C. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," in *Proceedings of ICLR*, 2017.

[11] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *Proceedings of ICASSP*. IEEE, 2018, pp. 4779–4783.

[12] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, pp. 3104–3112, 2014.

[13] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proceedings of ICASSP*. IEEE, 2019, pp. 3617–3621.

[14] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[15] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in neural information processing systems*, 2018, pp. 10 215–10 224.

[16] D. Erro, I. Sainz, I. Luengo, I. Odriozola, J. Sánchez, I. Saratxaga, E. Navas, and I. Hernáez, "HMM-based speech synthesis in Basque language using HTS," in *Proceedings of FALA*. RTTH, 2010, pp. 67–70.

[17] J. Wells, W. Barry, M. Grice, A. Fourcin, and D. Gibbon, "Standard computer-compatible transcription," *Esprit project 2589 (SAM), Doc. no. SAM-UCL*, vol. 37, 1992.

[18] T. Etchegoyhen, H. Arzelus, H. Gete, A. Alvarez, I. Hernaez, E. Navas, A. González-Docasal, J. Osácar, E. Benites, I. Ellakuria, E. Calonge, and M. Martin, "MINTZAI: Sistemas de Aprendizaje Profundo E2E para Traducción Automática del Habla MINTZAI: End-to-end Deep Learning for Speech Translation," *Sociedad Española para el Procesamiento del Lenguaje Natural*, vol. 65, pp. 97–100, 2020.

[19] Y. Ren, T. Qin, Y. Ruan, S. Zhao, T. Y. Liu, X. Tan, and Z. Zhao, "FastSpeech: Fast, robust and controllable text to speech," *arXiv*, may 2019.

[20] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 910–14 921.

[21] Y. A. Chung, Y. Wang, W. N. Hsu, Y. Zhang, and R. J. Skerry-Ryan, "Semi-supervised Training for Improving Data Efficiency in End-to-end Speech Synthesis," in *Proceedings of ICASSP*, vol. 2019-May. IEEE, 2019, pp. 6940–6944.

[22] NVIDIA, https://github.com/NVIDIA/tacotron2, 2018, online; accessed 20 December 2020.

[23] K. Ito and L. Johnson, "The LJ Speech Dataset, v1.1," https://keithito.com/LJ-Speech-Dataset/, 2017, online; accessed 20 December 2020.

[24] B. Külebi, A. A. Alp¨oktem, A. Peiró-Lilja, S. Pascual, and M. Farrús, "CATOTRON-A Neural Text-to-Speech System in Catalan," in *Proceedings of INTERSPEECH*. ISCA, 2020, pp. 490–491.

[25] X. Zhu, Y. Zhang, S. Yang, L. Xue, and L. Xie, "Pre-alignment guided attention for improving training efficiency and model stability in end-to-end speech synthesis," *IEEE Access*, vol. 7, pp. 65 955–65 964, 2019.

[26] S. Maiti and M. I. Mandel, "Speaker Independence of Neural Vocoders and Their Effect on Parametric Resynthesis Speech Enhancement," in *Proceedings of ICASSP*. IEEE, 2020, pp. 206–210.

[27] M. He, Y. Deng, and L. He, "Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS," in *Proceedings of INTERSPEECH*, vol. 2019-September. ISCA, jun 2019, pp. 1293–1297.

[28] E. Battenberg, R. J. Skerry-Ryan, S. Mariooryad, D. Stanton, D. Kao, M. Shannon, and T. Bagby, "Location-Relative Attention Mechanisms for Robust Long-Form Speech Synthesis," in *Proceedings of ICASSP*. IEEE, 2020, pp. 6194–6198.

[29] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Proceedings of INTERSPEECH*. ISCA, 2017, pp. 498–502.

[30] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," *Proceedings of INTERSPEECH*, pp. 1748–1752, 2020.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Multi-view Temporal Alignment for Non-parallel Articulatory-to-Acoustic Speech Synthesis

Jose A. Gonzalez-Lopez[1], *Miriam Gonzalez-Atienza*[1], *Alejandro Gomez-Alanis*[1], *José L. Pérez-Córdoba*[1], *and Phil D. Green*[2]

[1]University of Granada, Granada, Spain
[2]University of Sheffield, Sheffield, U.K.

joseangl@ugr.es

Abstract

Articulatory-to-acoustic (A2A) synthesis refers to the generation of audible speech from captured movement of the speech articulators. This technique has numerous applications, such as restoring oral communication to people who cannot longer speak due to illness or injury. Most successful techniques so far adopt a supervised learning framework, in which time-synchronous articulatory-and-speech recordings are used to train a supervised machine learning algorithm that can be used later to map articulator movements to speech. This, however, prevents the application of A2A techniques in cases where parallel data is unavailable, e.g., a person has already lost her/his voice and only articulatory data can be captured. In this work, we propose a solution to this problem based on the theory of multi-view learning. The proposed algorithm attempts to find an optimal temporal alignment between pairs of non-aligned articulatory-and-acoustic sequences with the same phonetic content by projecting them into a common latent space where both views are maximally correlated and then applying dynamic time warping. Several variants of this idea are discussed and explored. We show that the quality of speech generated in the non-aligned scenario is comparable to that obtained in the parallel scenario.

Index Terms: multi-view learning, dynamic time warping, canonical correlation analysis, silent speech interface, speech restoration.

1. Introduction

A silent speech interface (SSI) is a type of assistive technology aimed at restoring normal, vocal communication to speech-impaired persons. It does so by decoding speech from biosignals, different from the actual acoustic signal, generated by the human body during speech production. These biosignals can range from the neural activity in the speech and language areas of the brain [1–3], electrical activity driving the face muscles captured by surface electrodes (i.e., electromyography (EMG)) [4–6], or motion capture of the speech articulators by means of imaging techniques [7] or electromagnetic articulography techniques [8–11].

Because SSIs do not rely on the acoustic speech signal, they offer a radically new form of restoring oral communication to people with speech impairments. Two alternative SSI approaches have been proposed to decode speech from speech-related biosignals [12]: silent speech recognition, which involves the use of automatic speech recognition (ASR) algorithms to transform the biosignals into text, and direct speech synthesis, which directly maps the biosignals into a set of acoustic parameters amenable to speech synthesis. In this work, we focus on the latter approach, which is also known as articulatory-to-acoustic (A2A) synthesis in the literature when the biosignals encode information about the movements of the speech organs.

Most successful direct synthesis techniques so far adopt a data-driven framework in which supervised machine learning is used to model the mapping $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$ between source feature vectors $\boldsymbol{x}$ extracted from the biosignals and target feature vectors $\boldsymbol{y}$ computed from the speech signals. To train this function, a dataset $\mathcal{D} = \{(\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_M, \boldsymbol{Y}_M)\}$ with pairs of sequences of feature vectors $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ extracted from time-synchronous recordings is used, where $\boldsymbol{X}_i \in \mathbb{R}^{d_x \times T_i}$ and $\boldsymbol{Y}_i \in \mathbb{R}^{d_y \times T_i}$. The need for parallel recordings, however, limits the application of direct synthesis techniques to only a few clinical scenarios, as described in [12]. For instance, people who have already lost their voices could not use this technology because of the impossibility of recording parallel data.

A solution to this problem, which we explore in this work, involves the use of speech recordings from voice donors (e.g., relatives or recordings of the patient's own voice made before the voice loss). Using these recordings, it would be possible in principle to obtain the necessary parallel data by asking the patient to mouth along the speech recordings while articulatory data is captured. Even in this case, it is likely that the articulatory data will not be perfectly aligned with the speech recordings, having both modalities slightly different duration, thus preventing the direct application of standard supervised machine learning techniques.

To address this issue, in this work we propose an algorithm called multi-view temporal alignment by dependence maximisation in the latent space (TRANSIENCE)[1], which is based on the theory of multi-view learning [13, 14]. TRANSIENCE attempts to find the optimal temporal alignment between sequences of multiple views (e.g., audio and articulatory data) by first non-linearly projecting the data into a common, latent subspace where the views are maximally dependent and then aligning the resulting latent variables by means of the dynamic time warping (DTW) algorithm[2] [15]. We examine the performance of the proposed algorithm on a A2A task involving the conversion of articulatory data captured using permanent magnet articulography (PMA) [16, 17] to speech for multiple non-impaired

This work was funded by the Spanish State Research Agency (SRA) under the grant PID2019-108040RB-C22/SRA/10.13039/501100011033. Jose A. Gonzalez-Lopez holds a Juan de la Cierva-Incorporation Fellowship from the Spanish Ministry of Science, Innovation and Universities (IJCI-2017-32926).

[1]Code is available at https://github.com/joseangl/transience.

[2]It should be noted that the direct application of DTW to this problem is not possible because the views may have different dimensionality.

10.21437/IberSPEECH.2021-49

subjects.

The remainder of this paper is organised as follows. First, in Section 2, the relevant related work is reviewed and the main differences w.r.t. our technique are discussed. The details of the proposed technique are presented in Section 3. Section 4 describes the experimental setup, while the results are shown in Section 5. We conclude in Section 6 and discuss potential future research directions.

2. Related work

The most closely related work to our method is that of Trigeorgis *et al.* [18], in which an algorithm called deep canonical time warping (DCTW) combining canonical correlation analysis (CCA) [19] with DTW is proposed. The key differences of our approach w.r.t. DCTW can be summarised as follows. First, we evaluate different similarity metrics, not only CCA, to optimise the parameters of the deep neural networks (DNNs) used to map the multiple views into their common, latent subspace. Also, we introduce an autoenconder-based loss, which helps to regularise the training and avoids *naïve* solutions. Finally, inspired by the work in [20], we also propose the introduction of private latent variables for each view which aim at modelling the specific peculiarities within each view. Our technique also shares similarities with the generalyzed canonical time warping (GCTW) technique described in [21]. However, in contrast to this technique, TRANSIENCE solves the optimal alignment problem with DTW rather than approximating the temporal warping with a set of pre-defined monotonic bases and optimising the weights of these bases with a Gauss-Newton algorithm. Also, being based on CCA, GCTW computes the latent variables by *linearly projecting* the data from the different views, while our method uses powerful autoencoders to non-linearly transform the data, which we could be expected to yield better alignments.

Contrary to now popular sequence-to-sequence (seq2seq) models (e.g., [22, 23]), our method has different advantages. First, our aim is to align sequences of different lengths in order to obtain the necessary parallel data to train a machine learning technique, not to model a full-fledged mapping between sequences. In other words, our technique is only used to align the training data, while in test time the articulatory data is directly mapped to speech. Also, it is relatively straightforward to modify our method to process multiple views (more than 2), while the adaptation of seq2seq models is more involved.

3. Multi-view temporal alignment

The problem we address can be formulated as finding the optimal alignment between two time series $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_{T_x})$ and $\boldsymbol{Y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_{T_y})$ with possibly different dimensions. In our case, $\boldsymbol{X} \in \mathbb{R}^{d_x \times T_x}$ is a sequence of feature vectors extracted from an articulatory signal and $\boldsymbol{Y} \in \mathbb{R}^{d_y \times T_y}$ is the sequence of acoustic speech parameters. We further assume that both sequences encode the *same phonetic content* (i.e., same words in the same order) but have, possibly, different duration. Mathematically, this involves solving the following minimisation problem,

$$\underset{\boldsymbol{\phi}^x, \boldsymbol{\phi}^y}{\operatorname{argmin}} \sum_{t=1}^{T} d(\boldsymbol{x}_{\phi_t^x}, \boldsymbol{y}_{\phi_t^y}), \tag{1}$$

where $T = \max(T_x, T_y)$, $d(\boldsymbol{x}, \boldsymbol{y})$ is a distance function, $\boldsymbol{\phi}^x \in \{1 : T_x\}^T$ and $\boldsymbol{\phi}^y \in \{1 : T_y\}^T$ are warping functions that map the indices of the original time series to a common time axis. This way, $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ will be aligned if $\phi_t^x = i$ and $\phi_t^y = j$ for a given t.

To enable the alignment of sequences with different dimensionality, we assume that there exists a pair of transformation functions $\boldsymbol{f} : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_z}$ and $\boldsymbol{g} : \mathbb{R}^{d_y} \rightarrow \mathbb{R}^{d_z}$, modelled as DNNs in this work, that map the feature vectors into a common, latent space where the data from both views are maximally dependent (in the next section we will discuss how we measure this dependence). Thus, the problem in (1) now becomes,

$$\underset{\boldsymbol{\phi}^x, \boldsymbol{\phi}^y}{\operatorname{argmin}} \sum_{t=1}^{T} d(\boldsymbol{f}(\boldsymbol{x}_{\phi_t^x}), \boldsymbol{g}(\boldsymbol{y}_{\phi_t^y})). \tag{2}$$

For some fixed mapping functions $\boldsymbol{z}^x = \boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{z}^y = \boldsymbol{g}(\boldsymbol{y})$, the problem of temporal alignment of latent variable sequences $\boldsymbol{Z}^x = (\boldsymbol{z}_1^x, \dots, \boldsymbol{z}_{T_x}^x)$ and $\boldsymbol{Z}^y = (\boldsymbol{z}_1^y, \dots, \boldsymbol{z}_{T_y}^y)$ can be solved efficiently by means of the DTW algorithm [15]. Conversely, for fixed warping paths $\boldsymbol{\phi}^x$ and $\boldsymbol{\phi}^y$, the problem in (2) involves optimising the functions $\boldsymbol{f}(\cdot)$ and $\boldsymbol{g}(\cdot)$ in order to minimise $\sum_{t=1}^{T} d(\boldsymbol{f}(\boldsymbol{x}_{\phi_t^x}), \boldsymbol{g}(\boldsymbol{y}_{\phi_t^y}))$. If the mapping functions are modelled as DNNs, as in our case, this latter problem can be solved by back-propagation. Thus, TRANSIENCE algorithm solves (2) by alternating between the following two phases: (i) finding the optimum DNNs weights (Θ_f, Θ_g) by fixing the warping paths, and (2) applying DTW to compute the optimum warping paths (i.e., optimum alignments) (ϕ^x, ϕ^y) by freezing the DNNs weights. The warping paths are initialised by uniformly aligning the sequences, i.e., $\phi_t^x = 1 + \left\lceil \frac{t-1}{T-1}(T_x - 1) \right\rceil$ and $\phi_t^y = 1 + \left\lceil \frac{t-1}{T-1}(T_y - 1) \right\rceil$ for $t = 1, \dots, T$ and $T = \max(T_x, T_y)$.

3.1. Latent-space dependence metrics

A key component of our algorithm is the distance function $d(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{y}))$ used in (2) to evaluate the dependence between pairs of aligned latent variables. Here, we evaluate three alternative loss functions that optimise this dependence by maximising the correlation, mutual information and (minimise) a contrastive loss, respectively, between the latent variables.

3.1.1. Canonical correlation analysis

Given a mini-batch of N pairs of *aligned* observations $\mathcal{B} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \dots (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$, this loss function attempts to maximise the correlation between the outputs of the DNNs, $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{y})$, as follows,

$$(\Theta_f^*, \Theta_g^*) = \underset{\Theta_f, \Theta_g}{\operatorname{argmax}} \sum_{i=1}^{N} \operatorname{corr}(\boldsymbol{f}(\boldsymbol{x}_i; \Theta_f), \boldsymbol{g}(\boldsymbol{y}_i; \Theta_g)). \tag{3}$$

As detailed in [13], this equals to maximising the following loss function,

$$\mathcal{L}_{cca} = \sqrt{\operatorname{tr}(\boldsymbol{T}'\boldsymbol{T})}, \tag{4}$$

where $\operatorname{tr}(\cdot)$ is the trace operator and $\boldsymbol{T} = \boldsymbol{\Sigma}_{xx}^{-1/2} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1/2}$. The covariance matrices $\boldsymbol{\Sigma}_{xx} = \operatorname{cov}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}))$, $\boldsymbol{\Sigma}_{xy} = \operatorname{cov}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{y}))$ and $\boldsymbol{\Sigma}_{yy} = \operatorname{cov}(\boldsymbol{g}(\boldsymbol{y}), \boldsymbol{g}(\boldsymbol{y}))$ are estimated from the outputs of the DNNs. It should be noted that the DCTW algorithm described in [18] is a specific case of our TRANSIENCE algorithm when the CCA loss in (4) is used.

3.1.2. Maximum mutual information

As an alternative, we also consider maximising the mutual information between the latent variables as follows,

$$\mathcal{L}_{mmi} = \sum_{i=1}^{N} p(\boldsymbol{f}(\boldsymbol{x}_i), \boldsymbol{g}(\boldsymbol{y}_i)) \log \frac{p(\boldsymbol{f}(\boldsymbol{x}_i), \boldsymbol{g}(\boldsymbol{y}_i))}{p(\boldsymbol{f}(\boldsymbol{x}_i))p(\boldsymbol{g}(\boldsymbol{y}_i))}. \quad (5)$$

The probability density functions (pdfs) in (5) are estimated using kernel density estimation (KDE) as follows,

$$p(\boldsymbol{z}_i) = \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} K(\boldsymbol{z}_i - \boldsymbol{z}_j), \quad (6)$$

where an isotropic Gaussian kernel with trainable bandwidth σ_z is used in this work, i.e., $K(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \sigma_z \boldsymbol{I})$. Thus, three pdfs are estimated, the joint distribution $p(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{y}))$ and the marginals $p(\boldsymbol{f}(\boldsymbol{x}))$ and $p(\boldsymbol{g}(\boldsymbol{y}))$, each one with its own trainable bandwidth.

3.1.3. Contrastive loss

Finally, we also evaluate the contrastive loss function described in [20, 24] which, given a fixed latent variable from the first view $\boldsymbol{f}(\boldsymbol{x}^+)$, takes an aligned positive example $\boldsymbol{g}(\boldsymbol{y}^+)$ and an unaligned negative example $\boldsymbol{g}(\boldsymbol{y}^-)$ from the second view and attempts to *minimise* the difference between the distances for the positive and negative examples:

$$\mathcal{L}_{contrastive} = \frac{1}{N} \sum_{i=1}^{N} \max(0, m + d(\boldsymbol{f}(\boldsymbol{x}_i^+), \boldsymbol{g}(\boldsymbol{y}_i^+)) - d(\boldsymbol{f}(\boldsymbol{x}_i^+), \boldsymbol{g}(\boldsymbol{y}_i^-))), \quad (7)$$

where m is a margin hyperparameter ($m = 0.5$ is used in this work) and $d(\boldsymbol{z}_x, \boldsymbol{z}_y) = 1 - \frac{\boldsymbol{z}_x \cdot \boldsymbol{z}_y}{\|\boldsymbol{z}_x\| \|\boldsymbol{z}_y\|}$ is the cosine similarity. The negative examples $\boldsymbol{g}(\boldsymbol{y}^-)$ are generated by shuffling the outputs of the DNN for the second view before the loss is computed. Intuitively, the distances $d(\boldsymbol{f}(\boldsymbol{x}^+), \boldsymbol{g}(\boldsymbol{y}^+))$ in (7) should be small if both views are projected to similar (closer) representations in the common latent space, whereas the distances to the negative, unpaired examples $d(\boldsymbol{f}(\boldsymbol{x}^+), \boldsymbol{g}(\boldsymbol{y}^-))$ should be bigger because they are projected to different locations of that space.

3.2. Multi-view autoencoder

In order to regularise the training of the DNNs and to avoid naïve solutions (e.g., $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{y}) = \boldsymbol{c}$, for all $\boldsymbol{x}$ and $\boldsymbol{y}$, with $\boldsymbol{c}$ being a constant vector), we introduce an autoencoder-based reconstruction loss that minimises the mean squared error (MSE) between the DNNs' inputs $(\boldsymbol{x}, \boldsymbol{y})$ and the reconstructed outputs $(\hat{\boldsymbol{x}} = \boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x})), \hat{\boldsymbol{y}} = \boldsymbol{g}^{-1}(\boldsymbol{g}(\boldsymbol{y})))$, being $\boldsymbol{f}^{-1}$ and $\boldsymbol{g}^{-1}$ decoder networks that attempt to reconstruct $\boldsymbol{x}$ and $\boldsymbol{y}$ from their latent projections. The parameters of such networks, as well as those from the encoders $\boldsymbol{f}$ and $\boldsymbol{g}$, are trained by gradient-descend techniques using the following loss function,

$$\mathcal{L}_{autoencoder} = \frac{\lambda}{N} \left(\sum_{i=1}^{N} \left\| \boldsymbol{x}_i - \boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x}_i)) \right\|^2 + \sum_{i=1}^{n} \left\| \boldsymbol{y}_i - \boldsymbol{g}^{-1}(\boldsymbol{g}(\boldsymbol{y}_i)) \right\|^2 \right), \quad (8)$$

where the hyperparameter λ is set to 1 in this work.

Table 1: *Details of the dataset used for the experiments.*

Condition		# of sentences *Train*	# of sentences *Eval.*
Intra-subject	F → F	103 (11.3 min)	20 (1.2 min)
	M → M	134 (8.4 min)	20 (1.1 min)
Cross-subject	F → F	99 (6.0 min)	18 (0.9 min)
	F → M	332 (18.7 min)	20 (1.0 min)
	M → F	332 (21.9 min)	20 (1.2 min)
	M → M	414 (27.9 min)	20 (1.2 min)

3.3. Private latent variables

A key assumption of TRANSIENCE is that the views share some common information. Although this is indeed the case for audio and articulatory data, where both views encode the same phonetic information, it is also true that each view may have its own unique characteristics, thus making the reconstruction loss in (8) difficult to optimise when only considering the shared latent variables. For instance, the PMA technique used in this work for articulator motion capture is known to model poorly the information about the speech prosody [25, 26], whereas this information can be easily decoded from the acoustic signal. Thus, it may be beneficial to model the unique characteristics of each view as well as the common characteristics shared among all the views. Inspired by the work in [20], we propose the introduction of private latent variables for each view $\tilde{\boldsymbol{z}}^x$ and $\tilde{\boldsymbol{z}}^y$ that aim at modelling the uniqueness of each view. The private variables are predicted from the inputs by a set of independent DNNs $\tilde{\boldsymbol{z}}^x = \tilde{\boldsymbol{f}}(\boldsymbol{x})$ and $\tilde{\boldsymbol{z}}^y = \tilde{\boldsymbol{g}}(\boldsymbol{y})$. These private latent variables are used in (8), in addition to the common variables, for reconstructing the input data, i.e., $\hat{\boldsymbol{x}} = \boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x}), \tilde{\boldsymbol{f}}(\boldsymbol{x}))$ and $\hat{\boldsymbol{y}} = \boldsymbol{g}^{-1}(\boldsymbol{g}(\boldsymbol{y}), \tilde{\boldsymbol{g}}(\boldsymbol{y}))$.

When optimising the weights of DNNs $\tilde{\boldsymbol{f}}$ and $\tilde{\boldsymbol{g}}$, a standard Gaussian distribution $\tilde{\boldsymbol{z}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ is chosen as the prior distribution for the private variables of each view. To enforce this distribution, we minimise the Kullback-Leibler (KL) divergence between the priors and the empirical distribution (modelled as a multivariate Gaussian distribution with diagonal covariance) estimated from the private variables as follows,

$$\mathcal{L}_{KL} = \frac{1}{2} \sum_{i=1}^{d_{\tilde{z}}} \left(\sigma_i^2 + \mu_i^2 - 1 - \log \sigma_i^2 \right), \quad (9)$$

where the mean and variances in (9) are estimated for each private variable in each mini-batch.

4. Experimental setup

4.1. Dataset

The proposed alignment algorithm was evaluated on a A2A task involving the synthesis of speech from articulatory data captured using PMA [9, 17]. Parallel data was recorded by four non-impaired British subjects (2 males and 2 females) while reading aloud a subset of the CMU Arctic corpus [27]. Two alignment conditions were evaluated: (i) intra-subject alignment, where PMA and speech signals recorded by the same subject in different sessions are aligned, and (ii) cross-subject alignment, where PMA signals recorded by a given subject are aligned with speech recorded by a different subject (with possibly different gender as well). We also attempted to align PMA signals recorded from a female laryngectomy patient with a set of speech recordings made by her before losing the voice. The details of the dataset used for our experiments is summarised in Table 1.

Table 2: *Summary of the objective results.*

Method	MGCC MCD (dB)	BAP RMSE (dB)	F_0 RMSE (Hz)	Voicing err. rate(%)
Oracle	7.81	0.43	14.75	23.79
CTW	8.55	0.59	15.98	23.08
+autoenc.	9.20	0.88	16.70	25.30
+priv. vars.	8.83	0.58	15.74	**21.47**
CCA	9.37	0.85	16.40	27.95
+autoenc.	10.02	1.46	15.95	34.08
+priv. vars.	10.48	1.24	15.79	31.88
MMI	9.74	0.69	16.43	22.25
+autoenc.	9.97	1.09	16.92	23.72
+priv. vars.	9.86	1.70	16.41	21.75
Contrastive	**7.65**	**0.12**	15.28	24.10
+autoenc.	7.76	0.20	14.98	24.06
+priv. vars.	7.82	0.30	**14.58**	23.68

4.2. Implementation details

Each DNN in TRANSIENCE was modelled as a 3-layer feed-forward neural network with $200 \times 100 \times 100$ hidden units and leaky rectified linear unit (LReLU) activations ($a = 0.03$) following [18]. The neural networks were trained as denoising autoencoders ($\sigma_{noise} = 0.5$) using the Adam algorithm [28] with a fixed learning rate of $1e-4$ and a batch size of $N = 512$ samples. The dimensionality of the shared latent variables was set to $d_z = 20$ and fixed $d_{\tilde{z}} = 10$ for the private variables. The PMA signals were parameterised by applying principal component analysis (PCA) over contextual windows with 11 frames. The acoustic signals, on the other hand, were parameterised using the WORLD vocoder [29] with a frame rate of 5 ms as 25 mel-generalised cepstral coefficients (MGCCs), 1 band aperiodicity (BAP) value, 1 continuous F_0 value in logarithmic scale and 1 U/V decision. For temporal alignment, only the MGCCs were used (augmented with delta and acceleration parameters). Finally, the cosine distance was used for DTW.

4.3. PMA-to-speech system

The aligned signals were used to train speaker-dependent A2A systems. We used the same setting as in our previous work [26]: DNNs with 4 hidden layers with 400 units in each layer and rectified linear unit (ReLU) activations were used. The maximum-likelihood parameter generation (MLPG) algorithm [30,31] was applied over the DNN outputs to enhance the acoustic quality of the re-synthesised waveforms.

5. Results

5.1. Objective evaluation

First, we evaluated the quality of the re-synthesised speech signals obtained from the test PMA signals by comparing them with the original speech recordings made by the non-impaired subjects. For this task, we used several objective metrics widely used in speech synthesis: mel-cepstral distortion (MCD), root mean squared error (RMSE) of the predicted BAP and F_0 values and the error rate for the voicing parameter. For TRANSIENCE, three variants were evaluated depending on the latent-space similarity loss function employed: CCA, maximum mutual information (MMI), and Contrastive losses described in Section 3.1. Furthermore, for each of the three variants, we also evaluated the effect of the autoencoder-based loss described in Section 3.2 and the introduction of the private variables described in Section 3.3. For comparison purposes, we also evaluated the canonical time warping (CTW) technique described in [32], which is a particular case of TRANSIENCE combining standard (linear) CCA with DTW, thus not being able to model non-linear latent mappings. We also provide the results obtained by an oracle system, in which the PMA and acoustic signals in the training dataset are aligned by using the *ideal* warping paths computed by applying DTW over the speech signals recorded by the same subjects.

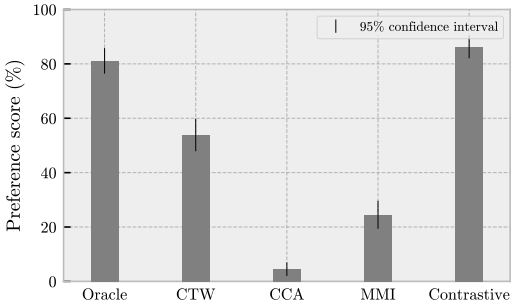


Figure 1: *Results of the ABX test on speech quality.*

Table 2 shows the objective results for the different systems. TRANSIENCE using the contrastive loss yields the best alignments and, hence, the best objective results, outperforming the other similarity metrics and, even, the oracle system. In particular, relative gains of 18.36% and 21.46% are achieved in the MCD metric w.r.t. using the CCA and MMI losses. Unfortunately, it seems that the introduction of the autoencoder-based loss and the private latent variables do not improve the results, which may be due to the dimensionality of the private variables not being enough to capture the peculiarities of each view. It is also surprising that the simple (linear) CTW technique outperforms its non-linear version TRANSIENCE+CCA. In future work, we should look at this issue.

5.2. Subjective evaluation

We also conducted an ABX test to subjectively evaluate the quality of the resynthesised speech signals. 27 listeners participated in the test, who have to judge which of two versions of the same signal produced by any combination of two of the 5 systems in Table 2 was more similar to a reference (one of the signals recorded by the subjects). Each listeners evaluated 10 sample pairs for each of the 10 possible system combinations (i.e., 100 pairs evaluated in total). For this task, only the "basic´´ systems in Table 2 were evaluated (i.e., without autoencoder-loss and private variables), because this setting produced the best objective results. Fig. 1 shows the results of the listening test. The most preferred system by a large margin was Contrastive, being on par with the Oracle system. Interestingly, the CTW system obtained higher preference scores than its non-linear version (TRANSIENCE+CCA) and that the MMI system. It may be due that the optimisation process get stuck in poor local-minima for the latter systems. However, more research is needed to shed some light into this problem.

6. Conclusions

We have proposed a new method for the alignment of times series with different dimensionality. Our evaluation on a A2A task involving the synthesis of audible speech from articulatory signals has shown that it is feasible to deploy direct synthesis techniques in non-parallel scenarios. Future work include evaluating our technique using more data from clinical population and introducing more constraints for the alignment (e.g., phonetic constraints).

7. References

[1] F. H. Guenther, J. S. Brumberg, E. J. Wright, A. Nieto-Castanon, J. A. Tourville, M. Panko, R. Law, S. A. Siebert, J. L. Bartels, D. S. Andreasen, P. Ehirim, H. Mao, and P. R. Kennedy, "A wireless brain-machine interface for real-time speech synthesis," *PLoS ONE*, vol. 4, no. 12, pp. e8218–e8218, 2009.

[2] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[3] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[4] T. Schultz and M. Wand, "Modeling coarticulation in EMG-based continuous speech recognition," *Speech Commun.*, vol. 52, no. 4, pp. 341—-353, 2010.

[5] M. Wand, M. Janke, and T. Schultz, "Tackling speaking mode varieties in EMG-based speech recognition," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 10, pp. 2515—2526, 2014.

[6] M. Janke and L. Diener, "EMG-to-speech: Direct generation of speech from facial electromyographic signals," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2375–2385, 2017.

[7] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Commun.*, vol. 52, no. 4, pp. 288–300, 2010.

[8] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain Lang.*, vol. 31, no. 1, pp. 26–35, 1987.

[9] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Med. Eng. Phys.*, vol. 30, no. 4, pp. 419–425, 2008.

[10] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "A silent speech system based on permanent magnet articulography and direct synthesis," *Comput. Speech. Lang.*, vol. 39, pp. 67–87, 2016.

[11] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Direct speech reconstruction from articulatory sensor data by machine learning," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2362–2374, 2017.

[12] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. Martín-Doñas, J. L. Pérez-Córdoba, and A. M. Gomez, "Silent speech interfaces for speech restoration: A review," *IEEE Access*, vol. 8, pp. 177 995–178 021, 2020.

[13] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. ICML*, 2013, pp. 1247–1255.

[14] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. ICML*, 2015, pp. 1083–1092.

[15] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[16] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," in *Proc. ICASSP*, 2018, pp. 5279–5283.

[17] J. M. Gilbert, S. I. Rybchenko, R. Hofe, S. R. Ell, M. J. Fagan, R. K. Moore, and P. Green, "Isolated word recognition of silent speech using magnetic implants and sensors," *Med. Eng. Phys.*, vol. 32, no. 10, pp. 1189–1197, 2010.

[18] G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "Deep canonical time warping for simultaneous alignment and representation learning of sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1128–1138, 2017.

[19] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 1936.

[20] W. Wang, X. Yan, H. Lee, and K. Livescu, "Deep variational canonical correlation analysis," *arXiv preprint arXiv:1610.03454*, 2016.

[21] F. Zhou and F. De la Torre, "Generalized canonical time warping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 279–294, 2015.

[22] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, 2014, pp. 3104–3112.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[24] K. M. Hermann and P. Blunsom, "Multilingual models for compositional distributed semantics," in *Proc. Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 58–68.

[25] J. A. Gonzalez, L. A. Cheah, J. Bai, S. R. Ell, J. M. Gilbert, R. K. Moore, and P. D. Green, "Analysis of phonetic similarity in a silent speech interface based on permanent magnetic articulography," in *Proc. Interspeech*, 2014, pp. 1018–1022.

[26] J. A. Gonzalez, L. A. Cheah, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Evaluation of a silent speech interface based on magnetic sensing and deep learning for a phonetically rich vocabulary," in *Proc. Interspeech*, 2017, pp. 3986–3990.

[27] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *5th ISCA Workshop on Speech Synthesis*, 2004, pp. 223–224.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] M. Morise, F. Yokomuri, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE T. Inf. Syst.*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[30] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp. 1315–1318.

[31] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.

[32] F. Zhou and F. Torre, "Canonical time warping for alignment of human behavior," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 22, 2009, pp. 2286–2294.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Generation of Synthetic Sign Language Sentences

Aitana Villaplana, Carlos-D. Martínez-Hinarejos

Escola Tècnica Superior d'Enginyeria Informàtica
Pattern Recognition and Human Language Technology research center
Universitat Politècnica de València, Camino de Vera, s/n, València, 46022, Spain
aivilmo@inf.upv.es, cmartine@dsic.upv.es

Abstract

Sign language is one of the most usual ways of communication for deaf people. Their inclusion in the society would be greatly improved if sign language can be easily used to communicate with other people that do not understand properly that language. Automatic recognition systems, based on machine learning techniques, could be very useful for this task, providing signers with tools that could be used to transcribe sign language into written language automatically. Many previous works have centered mainly in the recognition of single words, and different datasets of single words signs are available for estimating recognition models for this task. However, the recognition of whole sentences is difficult, since the acquisition of datasets of sentences is in general harder than the acquisition of single words. Thus, the possibility of generating sentences in sign language from single word datasets is very attractive to obtain automatic systems for decoding sign language sentences. In this work, we present an approximation for generating sign sentences from sign single words acquired by using the LeapMotion sensor. We study the different difficulties that presents this generation process. Results for real sign language sentences show that training with these synthetic sentences improves the decoding performance with respect to using only single words for training.

Index Terms: sign language recognition, human-computer interaction, data augmentation

1. Introduction

Sign language is a powerful communication tool for people with hearing difficulties, since it switches the communication from the audio channel to the visual channel. There is no universal sign language, and each linguistic zone defines its standard set of signs for communicating. Basically, sign language is based on hand and arms gestures, although other parts of the body (such as face expressions) are usually taken into account.

Automatic recognition of sign language is an important issue in order to include their users into society, since most people not pertaining to this group do not understand sign language. The automatic recognition could be based on machine learning techniques, which have demonstrated that they can successfully decode sign language into regular words in certain conditions [1, 2].

Most of the work made for sign language is performed on single word recognition [1, 3, 4]. In that case, it becomes a classification problem where each isolated sign must be assigned to a word in the corresponding vocabulary. However, when facing sentence recognition the problem becomes more difficult, since usually there is not available segmentation of the different signs. Following an approximation similar to that employed in regular speech recognition [5, 6] requires a considerable amount of data of sign language sentences from a given vocabulary. Contrarily to what happens with speech, there are only a few large sign language sentence corpora [2, 7] that allow to employ machine learning methods similar to those of speech recognition, which makes unfeasible developing systems that can solve this task.

Nevertheless, datasets for single word recognition are fairly more available and with sufficient data. Thus, it would be desirable to employ this data to train systems that can recognise sentences. Clearly, simple concatenation of the signs is not a correct solution, since in isolated words each sign starts and ends in a repose position that does not appear between the words in continuous sentences. Therefore, a correct generation of sentences from single words must remove this repose position. This is not the only task, and it is necessary to interpolate the intermediate positions between the end of one word in the sentence and the beginning of the next one (excluding the repose positions).

In this article, we present some techniques to automatically detect repose positions and to interpolate the intermediate positions between two consecutive words in a set of Spanish sign language words acquired by using the LeapMotion[1] sensor. The presented techniques are evaluated and the best option is used to generate a large amount of synthetic sign language sentences. The quality of the synthetic data is evaluated by using that data to train a sentence decoding system based on Hidden Markov Models (HMM) and to compare it with a system where only the isolated word signs are used for training. This evaluation is performed on a reduced set of real Spanish sign language sentences.

The article presents the following structure. Section 2 presents the relevant previous works on sign language recognition and the available datasets. Section 3 presents the used datasets, the different techniques for sentence generation, and the evaluation of the quality of the generation for the different alternatives. Section 4 presents the experimental framework and the results obtained with the system trained with the generated sentences and their comparison with the system trained with single words. Section 5 presents the conclusions and the future work lines.

2. State of the art

Automatic sign language recognition is a field that has been explored for a long time. Many initial approximations were developed to recognise a defined set of static signs, usually associated to letters and numbers. This is the case of [8], where video capture is used for getting the hand shape and Multilayer Perceptrons are used for recognising Colombian Sign Language. In the same fashion, there are a few Kaggle tasks that propose the same problem for American Sign Language, such like Sign

[1] https://www.leapmotion.com/product/desktop

10.21437/IberSPEECH.2021-50

Language MNIST[2] or ASL Alphabet[3]. With the LeapMotion sensor, a few works in this line have been developed [9, 10].

This static image recognition problem evolved to the recognition of single words, that imply hands movements and, consequently, a sequence of hands positions to be decoded. This is the case for American Sign Language [1, 11], German Sign Language [12], Chinese Sign Language [4], or Spanish Sign Language [13]. Most of the available resources have a limited vocabulary [11, 12, 13], although a few present a high number of words to be recognised [1, 4]. The acquisition of the hand gestures is done by different methods, being Kinect acquisition one of the most popular. The LeapMotion sensor was used as well for the acquisition of gestures in some works [13, 14].

The final step has been the recognition of sign language sentences. This problem is quite more difficult, since the acquisition of whole sentences becomes more difficult in terms of acquisition effort and conditions. Thus, the amount of datasets in this case is very few. One of the most popular is the RWTH-PHOENIX dataset [15, 16], that consists of a set of weather forecast videos with a vocabulary of more than 1000 words in German Sign Language. Another example is the CUNY corpus for American Sign Language [17], but this corpus was acquired with the purpose of developing animations of sign language. For Spanish Sign Language, a first small corpus (276 sentences, vocabulary of 65 words) was acquired by using the LeapMotion sensor and was made publicly available[4]. In the last years, a larger Spanish Sign Language corpus is being acquired [18]; this corpus contains both controlled condition acquisitions and TV weather forecast examples. The acquisition of this dataset is still in progress.

3. Data generation and evaluation

This section presents the sign language dataset employed in the experiments, along with the techniques employed for detecting the repose states and for connecting the different words that form a sentence.

3.1. Original dataset

The original dataset is the same that was employed in [13]. This dataset was acquired with the LeapMotion sensor. LeapMotion allows to obtain several points of the hands position. In our case, this dataset obtains the three-dimensional coordinates of each fingertip and the center of the palm, along with the angle in the different axis that forms the hand. Thus, for each hand we have a total of 21 features. Features are normalised and, consequently, their values range from -1 to 1 (except for angle values, that range from -3 to 3).

The dataset presents a vocabulary of 92 words. For isolated words, each word was acquired ten times for each one of four different signers. Thus, the total number of words is of 3680. For sentences, a single signer acquired 274 sentences from a reduced vocabulary (65 words), with lengths from 3 to 7 words.

The original dataset dealt with the signs performed with a single hand (right hand) by copying the values of that hand on the other (left hand). This was done to avoid errors when training the recognition models, since the absence of one hand is detected by LeapMotion as zero values, which leaded to very regular values that cannot be used for inferring Gaussian parameters. However, in our case it is necessary to keep these zero values (in order to perform a proper connection of the signs) but avoiding the associated data problem. We solved this situation by filling with zero values and adding some small Gaussian random noise (with $\sigma^2 = 0.01$).

In order to check the influence of this encoding change in the performance of the recognition system, we repeated the experiments presented in [13] for single word recognition with this new encoding. The best result with the original encoding provides an error rate of 10.6%, while with the new encoding the best obtained result is 10.5%. However, when repeating the sentence recognition with this new encoding, Word Error Rate (WER) results go from 11.8% to 16.4%, which makes us think that this new encoding is not initially suited for sentence recognition. Anyway, for our purpose of connecting single word signs to form a sentence, it is necessary to employ this new encoding.

3.2. Detection of repose states

The first step in the generation of the synthetic sentences is the elimination of the repose state at the end of the first word, and at the beginning of the last word, and at both for middle words. There is not a clear definition of what a repose state is, since defining them as the parts at the beginning or the end of the sign that have a zero value does not match with data. Thus, a definition of repose state must be given.

In our case, we based the repose state definition on the Euclidean distance between two consecutive vectors (frames) in the encoded sign sequence. Repose states are characterised for being at the beginning or the end of the word and by their slow (or null) movements. Thus, we can conclude that distance between two frames pertaining to the repose state is relatively small with respect to the distance between to frames pertaining to the real sign. This relative value can be calculated according to the maximum distance between any two consecutive frames in the whole sequence. Thus, we can define that consecutive frames at the beginning or at the end of the sequence whose distance is below a threshold of the maximum distance in the sequence pertain to the repose state.

The definition of this threshold would provide different lengths for the repose states. Moreover, the application of the threshold can be done in different manners. More specifically, we defined three different techniques:

1. Fixed threshold: the chosen threshold is not changed when applied to the sequences at the beginning or at the end of the sign sequence.
2. Dual variable threshold: when the application of the chosen threshold provides no repose states (zero length at the beginning and at the end), the threshold is increased and the repose states are recalculated; the process is repeated until no lack of repose is obtained or a maximum threshold is applied.
3. Initial and final variable threshold: similar to the previous one but applied when any repose state is zero length and only on the parts that present that zero length.

We evaluated the performance of the different alternatives by calculating the percent of words that presented an inappropriate repose state. A repose state is considered inappropriate when is zero length (lack of repose) or is more than 30% of the

[2] https://www.kaggle.com/datamunge/sign-language-mnist/metadata

[3] https://www.kaggle.com/grassknoted/asl-alphabet

[4] https://github.com/zparcheta/spanish-sign-language-db

Table 1: *Percent of the words with inappropriate repose state for different values of the fixed threshold.*

Threshold (%)	Inappropriate words (%)
5	77.2
10	73.9
15	73.0
20	73.7

Table 2: *Percent of the words with inappropriate repose state for variable threshold, for values 5-15 and 5-20 of the threshold.*

	Inappropriate words (%)	
Threshold interval (%)	Dual	Initial/final
5-15	56.4	60.2
5-20	50.3	55.9

total sequence length or the sum of the beginning and the end repose is more than a 60% of the total (excess of repose). These criteria are based on empirical observation of the signs.

For the fixed threshold, we employed values ranging from 5% to 20% in steps of 5. The percent of the words that obtained an inappropriate repose is presented in Table 1. These results show that the method is not very effective; a detailed analysis showed that this method causes lack of repose in many cases, specially for the 5% threshold, which is the main source of errors. Thus, variable threshold is expected to improve this situation.

Results for dual variable threshold and initial and final variable threshold are presented in Table 2. Results show an improvement in the calculations of the repose states, specially for the dual technique until a 20% of threshold.

3.3. Interpolation between consecutive signs

Once the repose states are detected and removed from the sign sequence, consecutive words must be concatenated by interpolating a set of points that simulates the transition between one sign and the next one.

The approximation we followed was using trace segmentation [19] between the final point of one word and the initial point of the next word in the sentence to be generated. Trace segmentation infers a linear route between the two points that can be used to interpolate intermediate points.

One decision to be taken is how many points are going to be interpolated. In order to have a proper estimation of this number of points, it is necessary to compute how do hands usually progress in the generation of sentences, in particular between consecutive words. Therefore, distances between the final point of one word and the initial point of the next word were calculated, giving an average value of 8.35. Thus, synthetic sentences should present a similar difference.

Table 3: *Recognition results (WER) for 274 synthetic sentences generated with different repose detection and a fixed number of points for interpolation. For all percents of fixed and variable initial-final repose detection the results were the same.*

		# interpolation points			
Repose detection		3	4	5	6
Fixed	5%-20%	6.5	6.6	6.6	6.7
	5-15%	6.5	6.7	6.7	6.8
Dual	5-20%	6.6	6.8	6.8	6.9
Initial-Final	5-15%/20%	6.6	6.7	6.7	6.8

Table 4: *Recognition results (WER) for 274 synthetic sentences generated with different repose detection and a variable number of points for interpolation. For all percents of fixed and variable initial-final repose detection the results were the same.*

Repose detection		WER
Fixed	5%-20%	6.5
	5-15%	6.5
Dual	5-20%	6.6
Initial-Final	5-15%/20%	6.5

Table 5: *Recognition results (WER) with different noise factors for synthetic sentences with fixed repose detection (10% threshold) and variable number of point interpolation.*

Noise	WER
20%	7.0
30%	8.2
40%	11.5
50%	14.7
55%	16.6
60%	20.3
65%	21.8
70%	24.5

Initial tests showed that a number of points between 3 and 6 kept similar values for synthetic sentences. After that, the specific value for the selected number of points is calculated by using equidistant points in the linear plane defined by trace segmentation. The number of points can be fixed for all word combinations or can be variable according to the distance between the final point of one word and the initial point of the next word in the synthetic sentence (the more the distance, the more the number of points). In order to avoid a completely linear route between the connected points, some random noise can be introduced in the calculated points. Thus, it is necessary to introduce noise in order to obtain synthetic sentences

3.4. Selection of the repose detection and number of points

In order to select the final values for the repose selection technique (fixed, dual variable, initial and final variable) and the number of points for interpolation (fixed or variable), an experiment was performed with a set of 274 synthetic sentences that contained the same word sequences than their counterparts in the real sentences dataset. This set was generated for many different combinations of repose detection and interpolation.

The generated sentences were used in a four-fold cross-validation approach similar to that used in [13], where an HMM-based system (each word an HMM) was trained with all the single word samples and three partitions of the sentences, using the remaining sentences for test. Training of the HMMs was done in HTK [20] following the variable topology for each word described in [13] (with factor 1, which is the one that gave the best results). After initial tests, HMMs with 2 gaussians per state were used as those that offered a better performance.

The results obtained for the different number of fixed points are presented in Table 3. The results obtained for the variable number of points are presented in Table 4. These results show that differences among using any option are minimal, and that the synthetic sentences fit too much to the isolated words (WER

Table 6: *Real sign sentence recognition results (WER) with HMMs trained with only isolated words and isolated words plus synthetic sentences, for different noise factors and Gaussian number. Best results in boldface. Confidence intervals are in all cases lower than 3.6.*

Gaussians	Isolated words (baseline)	Isolated word and synthetic sentences								
		Noise factor (%)								
		0	10	20	30	40	50	55	60	65
1	52.2	36.9	37.7	37.0	36.0	35.8	34.8	35.0	**34.4**	**34.4**
2	52.2	36.7	38.3	36.6	36.2	36.1	34.6	35.2	35.7	35.3
4	52.0	36.9	38.1	36.9	36.4	36.1	35.1	35.0	36.7	35.3
8	52.0	37.0	38.2	37.0	36.4	36.4	35.1	35.0	36.3	34.9

with real sentences is about 16.4%).

Therefore, it is necessary to introduce some noise in order to obtain synthetic sentences whose behaviour is similar to the real sentences. Initially, noise was only introduced into the interpolation points, as described in Subsection 3.3, but given the small number of interpolated points with respect to real points, results barely changed. Thus, it was decided to apply a noise factor to all the vectors of the synthetic sentences. Taking as baseline system the one with repose detection by fixed threshold, with 10% threshold, and variable number of points in interpolation, several recognition experiments with different noise factors were performed, and their results are shown in Table 5. As it can be seen, a noise factor of 55% provides synthetic sentences with similar behaviour (in WER terms) to the real sentences.

As a conclusion, we can take the synthetic generation system with fixed threshold, 10% threshold, variable number of points in interpolation, and noise injection of 55% as an appropriate system for generating synthetic sentences that can be used to train models that improve the recognition of real sentences. This is the objective presented in Section 4.

4. Experiments and results

In this section, the use of synthetic sign language sentences to train HMMs in order to recognise real sign language sentences is studied. In general, the generated synthetic sentences are employed as supplementary training data, along with the isolated words, to train HMMs by using the HTK toolkit. The usual process consists of the steps of defining an initial bare HMM for each word, with one Gaussian per state as emission distribution, perform several Baum-Welch training iterations with all available data (words and synthetic sentences), and perform Gaussian increments.

The set of synthetic sentences that are generated are based on the transcription of the real sentences provided by the corpus we described above. For each transcription, about 50 different combinations of the different words repetitions available were generated (with the only restriction of keeping the same signer in the combined words). The final number of generated sentences is 13677 (some combinations appeared repeatedly and consequently there are not exactly 50 for each sentence transcription). They were generated with the parameters stated in Subsection 3.4: fixed repose detection with threshold 10%, variable number of points in interpolation. Noise factor is one of the parameters that is studied in these experiments.

Experiments consisted of the recognition of the real sign language sentences with different HMMs sets: those trained only with original isolated words without any repose removing (baseline) and those trained with both isolated words and synthetic sentences (with different noise factors). The experiments were performed for different number of Gaussians. The language model is the same used in [13]. Final results are presented in Table 6. Confidence intervals were calculated using bootstrapping [21].

As it can be seen, introducing synthetic sentences in the training process causes a high and significant increment in performance with respect to using only isolated words. Best results are obtained with a low number of Gaussians, which is reasonable given that data variability is not very high because of its synthetic nature. With respect to noise injection, best results are obtained with a noise injection of 60%/65% (very close to the optimal value of 55% determined in Section 3.4), although these results are not significantly better than other combinations.

In conclusion, the introduction of synthetic sentences in the training process seems to cause an increment in the recognition performance. However, this impact is quite far from that obtained when using real sentences (in that case, recognition results present a 16.4 WER). Therefore, more sophisticated techniques must be used to improve the representativity of the generated synthetic sentences.

5. Conclusions and future work

The use of synthetic generation of sign language sentences from isolated word signs could be an important source of complementary data to improve recognition performance of machine learning based methods. The study we have presented showed the feasibility of this generation and that the addition of these synthetic sentences is beneficial from the recognition performance point of view.

However, the synthetic sentences are still far from providing the same performance than the systems that employ real sentences for training. Thus, future work will be directed to improve the quality of the synthesis according to its proximity to the real sentences that are available. The use of techniques based on speech synthesis, such as HMM/DNN-based Speech Synthesis System (HTS) [22], could be an option for a better synthetic generation. Apart from that, current experiments have been only performed with HMMs as a prove of concept, but it would be desirable to exploit this synthesis to generate massive data that can be employed in the use of deep learning methods for sign language recognition.

6. Acknowledgements

This work was partially supported by Generalitat Valenciana under project DeepPattern (PROMETEO/2019/121) and by Ministerio de Ciencia under project MIRANDA-DocTIUM (RTI2018-095645-B-C22).

7. References

[1] M. Dilsizian, P. Yanovich, S. Wang, C. Neidle, and D. N. Metaxas, "A new framework for sign language recognition based on 3d handshape identification and linguistic modeling," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. European Language Resources Association (ELRA), 2014, pp. 1924–1929. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/summaries/1138.html

[2] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep sign: Enabling robust statistical continuous sign language recognition via hybrid cnn-hmms," *Int. J. Comput. Vis.*, vol. 126, no. 12, pp. 1311–1325, 2018. [Online]. Available: https://doi.org/10.1007/s11263-018-1121-3

[3] A. M. Martínez, R. B. Wilbur, R. Shay, and A. C. Kak, "Purdue rvl-slll asl database for automatic recognition of american sign language." in *ICMI*. IEEE Computer Society, 2002, pp. 167–172.

[4] H. Wang, X. Chai, X. Hong, G. Zhao, and X. Chen, "Isolated sign language recognition with grassmann covariance matrices," *ACM Trans. Access. Comput.*, vol. 8, no. 4, pp. 14:1–14:21, 2016. [Online]. Available: https://doi.org/10.1145/2897735

[5] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge, MA, USA: MIT Press, 1997.

[6] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19 143–19 165, 2019.

[7] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, pp. 108 – 125, 2015.

[8] J. D. Guerrero-Balaguera and W. J. Pérez-Holguín, "FPGA-based translation system from colombian sign language to text," *DYNA*, vol. 82, pp. 172 – 181, 2015.

[9] D. Naglot and M. Kulkarni, "Real time sign language recognition using the leap motion controller," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, vol. 3, 2016, pp. 1–5.

[10] T.-W. Chong and B.-G. Lee, "American sign language recognition using leap motion controller with machine learning approach," *Sensors*, vol. 18, no. 10, p. 3554, Oct 2018. [Online]. Available: http://dx.doi.org/10.3390/s18103554

[11] C. Chen, B. Zhang, Z. Hou, J. Jiang, M. Liu, and Y. Yang, "Action recognition from depth sequences using weighted fusion of 2d and 3d auto-correlation of gradients features," *Multim. Tools Appl.*, vol. 76, no. 3, pp. 4651–4669, 2017. [Online]. Available: https://doi.org/10.1007/s11042-016-3284-7

[12] E. Ong, O. Koller, N. Pugeault, and R. Bowden, "Sign spotting using hierarchical sequential patterns with temporal intervals," in *Proceeedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2014)*. IEEE, 2014, pp. 1931–1938.

[13] C. D. Martínez-Hinarejos and Z. Parcheta, "Spanish sign language recognition with different topology hidden markov models," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*. ISCA, 2017, pp. 3349–3353. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0275.html

[14] J. J. Bird, A. Ekárt, and D. R. Faria, "British sign language recognition via late fusion of computer vision and leap motion with transfer learning to american sign language," *Sensors*, vol. 20, no. 18, p. 5151, 2020. [Online]. Available: https://doi.org/10.3390/s20185151

[15] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, and H. Ney, "RWTH-PHOENIX-weather: A large vocabulary sign language recognition and translation corpus," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 3785–3789. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/844_Paper.pdf

[16] J. Forster, C. Schmidt, O. Koller, M. Bellgardt, and H. Ney, "Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 1911–1916. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/585_Paper.pdf

[17] P. Lu and M. Huenerfauth, "Collecting and evaluating the cuny asl corpus for research on american sign language animation," *Computer Speech & Language*, vol. 28, no. 3, pp. 812 – 831, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230813000879

[18] L. Docío-Fernández, J. L. Alba-Castro, S. Torres-Guijarro, E. Rodríguez-Banga, M. Rey-Area, A. Pérez-Pérez, S. Rico-Alonso, and C. García-Mateo, "LSE_UVIGO: A multi-source database for Spanish Sign Language recognition," in *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 45–52. [Online]. Available: https://www.aclweb.org/anthology/2020.signlang-1.8

[19] M. Kuhn, H. Tomaschewski, and H. Ney, "Fast nonlinear time alignment for isolated word recognition," in *ICASSP '81. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 6, 1981, pp. 736–740.

[20] S. Young, G. Evermann, M. Gales, T. Hain, D. K. aw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. V. hev, and P. C. Woodland, *The HTK book*. Cambridge university engineering department, 2006.

[21] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performanc e evaluation," in *Proc. of ICASSP*, vol. 1, 2004, pp. 409–412.

[22] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, B. AW, and K. Tokuda, "The hmm-based speech synthesis system version 2.0," in *Proceedings of ISCA Speech Synthesis Workshop 6*, 2007, pp. 131–136.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Contribution of vocal tract and glottal source spectral cues in the generation of happy and aggressive [a] vowels

Marc Freixes, Francesc Alías and Joan Claudi Socoró

GTM – Grup de recerca en Tecnologies Mèdia, La Salle - Universitat Ramon Llull
Quatre Camins, 30, 08022 Barcelona, Spain
{marc.freixes,francesc.alias,joanclaudi.socoro}@salle.url.edu

Abstract

At present, three-dimensional (3D) acoustic models allow for the numerical simulation of vowels, diphthongs and some vowel-consonant-vowel sequences using realistic vocal tract geometries. While research is being done to generate more phonemes and short utterances, some attempts have been made to incorporate expressiveness into the 3D numerical simulation of isolated vowels. However they are very preliminary and still far from the generation of expressive utterances. To move towards this goal, this work analyses the contribution of vocal tract (VT) and glottal source spectral (GSS) cues to the production of happy and aggressive vowels with respect to neutral vowels. After parameterising with the GlottDNN vocoder the paired neutral-expressive utterances from a Spanish database, neutral utterances are transplanted with the target expressive prosody as baseline, and subsequently resynthesised considering also the GSS and/or VT from their expressive pairs. Objective and subjective evaluations show that, both GSS and VT have a statistically significant contribution to convey the tense voice target emotions. VT prevails over GSS specially for aggressive. Best results are achieved when considering both GSS and VT, which compared to the baseline permits an increase in the perceived emotional intensity of 55.3% for happy and 62.8% for aggressive utterances.

Index Terms: expressive speech synthesis, emotional corpora, speech analysis, inverse filtering, glottal source, vocal tract, numerical voice production

1. Introduction

Speech synthesis methods that rely on the acoustic theory of voice production have traditionally considered simplified source-filter models [1], such as the ones based on one-dimensional (1D) representations of the vocal tract [2]. Recently, the increase in computing power has allowed the development of three-dimensional (3D) acoustic models, which have been applied to generate vowels [3], diphtongs [4] and some vowel-consonant-vowel sequences [5], overcoming the limitations of 1D-based models [6]. While researchers keep investigating on the production of other phonemes and short utterances [7], preliminary attempts to synthesise expressive vowels have been made, but limited to basic modifications of glottal source signals [8].

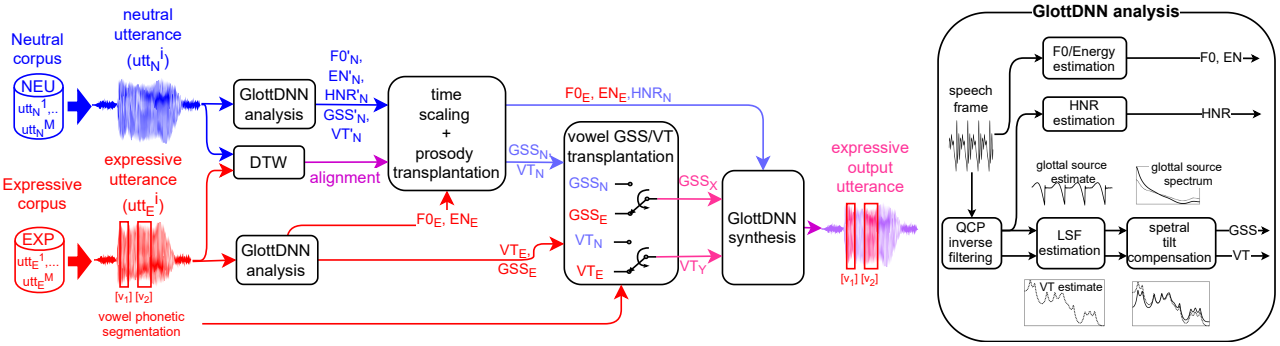
The relevance of glottal source and/or vocal tract cues in the production of expressive speech has been explored in several studies using traditional source-filter based approaches on an analysis-by-synthesis scheme. For example, the contribution of phonation types to the perception of emotions was analysed in [9]. To this end, a set of utterances was resynthesised with different phonation types through an articulatory-based synthesiser that incorporates a self-oscillating model of the vocal folds. A similar approach was followed in [10] considering a formant-based synthesiser with a modified Liljencrants-Fant (LF) glottal flow model [11]. This synthesis approach was also considered in [12] to study the mapping of F0 contours and voice quality on affect for different languages by modifying the parameters of modal stimuli. In [13], the LF model was controlled by modifying the R_d glottal shape parameter [14] to simulate the tense-lax continuum and explore its influence on emotion perception. Similarly, an auto-regressive exogenous LF model was proposed in [15] to analyse the contribution of glottal source and vocal tract to the perception of emotions in a valence-arousal space. Nevertheless, the study only evaluated isolated vowels, and suffered from the considered prosody *neutralisation* process.

The aforementioned approaches have focused on the analysis and resynthesis of a small set of vowels or utterances, some of them involving costly manual tuning processes. Nonetheless, recent advances in inverse filtering and glottal source processing techniques have facilitated the automatic decomposition of the speech signal into glottal source and vocal tract features [16]. For instance, GFM-Voc (Glottal Flow Model-based Vocoder) allows real-time voice manipulations, such as vowel formants shifting and voice quality modifications related to the glottal source [17]. Also included in this strand are glottal vocoders like GlottHMM, whose features proved effective in the analysis of expressive nuances in [18]. More recently, its successor, GlottDNN [19], was used to perform speaking style conversion to mimic the Lombard effect from natural speech [20]. A GlottDNN-based analysis of the glottal source spectral tilt was proposed [21] to introduce expressiveness in the 3D numerical generation of isolated vowels through modifications of the R_d parameter of the LF model. However, this proof-of-concept is still far from generating natural expressive utterances.

As a next step towards this goal, in this work we analyse the contribution of vocal tract (VT) and glottal source spectral (GSS) cues in the generation of emotional styles with tense phonation. For this purpose, paired neutral, happy and aggressive utterances from a Spanish speech database are inverse filtered and parameterised using the GlottDNN vocoder. Then, neutral utterances transplanted with prosody, and GSS and/or VT from the expressive pairs are resynthesised and evaluated through both objective and subjective tests, focused on vowels [a]; the most common vowel in the database.

The paper is organised as follows. Section 2 presents the methodology proposed for the GlottDNN-based analysis and synthesis of expressive utterances to study the contribution of GSS and VT on tense voice emotional styles. Next, the conducted experiments are described and the obtained results discussed in Sections 3 and 4, respectively. Finally, conclusions and future work are presented in Section 5.

10.21437/IberSPEECH.2021-51



(a) *Overall diagram.* (b) *Detail of the GlottDNN analysis.*

Figure 1: *Framework proposed to study the contribution of vocal tract and glottal source spectral cues (VT and GSS) in the generation of tense voice expressive (EXP) vowels, depicting the overall diagram of the process in (a) and the main elements of the GlottDNN analysis in (b). M pairs of neutral and expressive utterances from parallel speech corpora are analysed using the GlottDNN vocoder, and aligned through dynamic time warping (DTW). According to this alignment, the features of each neutral utterance utt_N^i (marked with $'$) are time-scaled and transplanted with the prosody of the corresponding expressive pair utt_E^i ($F0_E$, EN_E). Finally, different expressive output utterances are obtained by applying GSS_X and VT_Y to the vowels –being X, Y either neutral (N) or expressive (E).*

2. Methodology

Figure 1 depicts the framework proposed to analyse the contribution of the spectral cues from glottal source and vocal tract to the synthesis of tense voice expressive styles with respect to parallel neutral speech data. The study takes the neutral utterances with the prosody transplanted from their expressive pairs as the baseline. The contribution of GSS and VT is analysed through different synthesis configurations, which are denoted as GSS_XVT_Y, where X and Y indicate the origin of the GSS and VT applied to the vowels: N for the neutral utterance and E for the expressive utterance (hereafter, this subindex notation is applied for all the variables).

Firstly, each neutral-expressive utterance pair is parameterised using the GlottDNN vocoder. As depicted in Figure 1b, for each input speech frame (either neutral or expressive), the GlottDNN estimates its fundamental frequency (F0 in Hz) and energy (EN in dB), and applies quasi-closed phase (QCP) inverse filtering to obtain the corresponding glottal source and VT estimates, which are parameterised using Line Spectral Frequencies (LSF). Moreover, it computes the Harmonic-to-Noise Ratio (HNR) of the glottal source estimate. Further details of this process can be found in [19]. Next, a spectral tilt compensation module is included to compensate the tendency of QCP to include residual spectral cues from the glottal source on the vocal tract estimation [20]. In this module, the spectral tilt of the vocal tract estimate is modelled with a first order linear prediction filter, and subsequently transferred to the GSS to adjust it following [20].

Secondly, the prosody of the neutral utterance is transplanted by that from the expressive pair. On one hand, the GlottDNN features of the neutral utterance (marked with $'$ on Figure 1a) are linearly interpolated to time-scale them to the expressive target according to the alignment obtained through dynamic time warping. On the other hand, F0_N and EN_N are replaced by those from the expressive utterance, that is, F0_E and EN_E, respectively. Next, GSS_E and/or VT_E are transplanted into the vowels depending on the selected synthesis configuration.

The GlottDNN synthesis process is briefly described below (for further details the reader is referred to [19]). The GlottDNN vocoder uses white noise sequences as input to generate the unvoiced frames. Conversely, the vocoder features of the voiced frames are input into a simple feed-forward deep neural network to generate zero-padded two pitch-period glottal flow derivative pulses. These signals are scaled according to EN_E, and concatenated through the classic Pitch-Synchronous Overlap and Add (PSOLA) algorithm [22]. This initial excitation is then processed adding noise in the spectral domain as specified by the HNR_N besides modifying its spectra to match GSS_X. Finally, the excitation signal is filtered according to the considered VT_Y to obtain the synthetic speech output. It should be noted that since this work focuses on the spectral characteristics of the glottal source (i.e., GSS), the synthesis has been performed considering HNR_N and using the GlottDNN pulse generation model trained with neutral speech. The relevance of these two aspects of the glottal source in the generation of expressive speech will be analysed in future studies.

3. Experiments

This section describes the conducted experiments, detailing the main characteristics of the emotional speech database and the configuration of the GlottDNN vocoder, together with the design of the objective and subjective evaluations.

3.1. Emotional speech database

The experiments have been conducted on an emotional Spanish speech database explicitly designed to elicit expressive speech and recorded by a female professional speaker at a sampling frequency of 16 kHz [23].

Three out of the five expressive styles of that database have been chosen as the ones characterised by a modal or a tense phonation, namely: (i) neutral; (ii) happy; (iii) and aggressive. These corpora count with a set of 1250 paired short utterances (with an average of 1.2 words per utterance) that ensure phonetic coverage for Spanish text-to-speech synthesis purposes. Out of them, 841 utterances have been used in this work, specifically those containing at least a vowel [a]; the most common vowel in the database. As a result, a total of 1171 paired vowels have been considered in the experiments.

Table 1: *Mean values of the spectral distances (either Itakura-Saito, or $\overline{d_{\mathrm{IS}}}$, and Kullback-Leibler, or $\overline{d_{\mathrm{KL}}}$) computed from the analysed configurations to the expressive configurations for happy and aggressive [a] vowels.*

	Happy	Aggressive
$\overline{d_{\mathrm{IS}}}(\mathrm{GSS_N}, \mathrm{GSS_E})$	0.14	0.08
$\overline{d_{\mathrm{IS}}}(\mathrm{VT_N}, \mathrm{VT_E})$	3.11	3.79
$\overline{d_{\mathrm{KL}}}(\mathrm{GSS_N VT_N}, \mathrm{GSS_E VT_E})$	2.27	2.42
$\overline{d_{\mathrm{KL}}}(\mathrm{GSS_E VT_N}, \mathrm{GSS_E VT_E})$	1.17	1.90
$\overline{d_{\mathrm{KL}}}(\mathrm{GSS_N VT_E}, \mathrm{GSS_E VT_E})$	0.45	0.17

3.2. GlottDNN-based analysis and synthesis

The analysis and synthesis of utterances have been done using the default GlottDNN settings[1], parameterising GSS and VT with 10 and 30 LSF coefficients per frame, respectively, and considering voiced frames of 25 ms, and unvoiced frames of 15 ms. The whole neutral corpus (of 2.4 h length) has been used to train the GlottDNN pulse generation model [19].

3.3. Objective and subjective evaluation

The objective contribution of GSS and VT on the generation of happy and aggressive emotions has been evaluated through the computation of spectral distances between the [a] vowel pairs taking the expressive vowel as the target reference.

Given a neutral-expressive vowel pair, 2 GSSs and 2 VTs are obtained: from the neutral vowel ($\mathrm{GSS_N}$, $\mathrm{VT_N}$) and from the expressive vowel ($\mathrm{GSS_E}$, $\mathrm{VT_E}$). The similarity between the $\mathrm{GSS_N}$ and the $\mathrm{GSS_E}$ is computed as the Itakura-Saito LPC-based spectral distance, i.e., $d_{\mathrm{IS}}(\mathrm{GSS_N}, \mathrm{GSS_E})$. The same is done for the VT, i.e., $d_{\mathrm{IS}}(\mathrm{VT_N}, \mathrm{VT_E})$. GSS and VT are parameterised by the GlottDNN as LSF vectors at a frame level, from which LSF vectors at vowel level are obtained using the median to reduce coarticulation effects. Finally, LSF are translated into LPC to compute the Itakura-Saito distances [24].

Regarding the synthesis, a total of 4 configurations per emotion have been considered to evaluate the contribution of GSS and VT to the production of the tense voice target emotion, considering in all of them the target expressive prosody: (i) $\mathrm{GSS_N VT_N}$ as the baseline configuration; (ii) $\mathrm{GSS_E VT_N}$; (iii) $\mathrm{GSS_N VT_E}$; and (iv) $\mathrm{GSS_E VT_E}$ as the expressive target configuration. In order to evaluate how close is each vowel version to the expressive target, their long term average spectrum (LTAS) have been computed as the Welch's power spectral estimate, with a 15 ms hamming window, 50% of overlap and a 2048-point FFT [8]. Then, the similarity of each vowel obtained from configurations (i) to (iii) with the expressive target has been measured as the symmetrical Kullback-Leibler spectral distance [25] between its LTAS and the corresponding one in configuration (iv), i.e., $d_{\mathrm{KL}}(\mathrm{GSS}_X \mathrm{VT}_Y, \mathrm{GSS_E VT_E})$.

Regarding the subjective evaluation, the perceived emotional intensity for the different synthesis configurations has been assessed through a MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) perceptual test [26]. Six words from the speech database subset containing only [a] vowels were used to evaluate the contribution of GSS and VT to the perception of the two target emotions through the four aforementioned configurations. The words included in the test are *pala*, *taza*, *capa*, *gastar*, *agravar* and *MACBA* (in English, they correspond to shovel, cup, layer, spend, aggravate and MACBA –the name of a museum in Barcelona). Two versions of the test were prepared, each one consisting of 7 evaluation sets (three words per emotion plus one control point to validate the evaluator consistency according to the Pearson correlation coefficient r). In each set, the participants were asked to rate the perceived emotion intensity for each one of the four versions of the word on a 0 to 100 scale. A GlottDNN-based resynthesised utterance different from those evaluated was included in the test as an example of the target happy/aggressive emotion. In order to determine if the differences between the evaluated configurations are statistically significant, the Wilcoxon signed-rank test [27] has been applied to both objective and subjective results.

Forty-four Spanish native speakers with an average age of 28.9 took one of the two versions of the online test using headphones and the Web Audio Evaluation Tool [28]. Among them, 61.9% of the participants are engineering students in their final year, 42.9% have experience in playing and/or producing music, 21.4% in audio software/hardware design and the 28.6% in audio or speech research. Once the perceptual test was concluded, the responses of eight participants were discarded since they presented significant criteria inconsistencies (i.e., with $r < 0.5$).

[1] https://github.com/ljuvela/GlottDNN

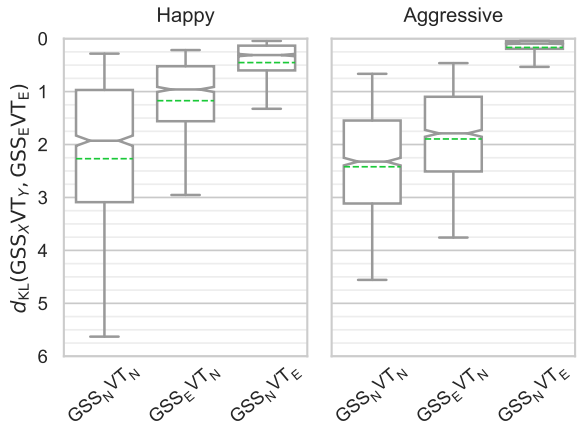


Figure 2: *Boxplots of the Kullback-Leibler distances from the analysed configurations to the target configuration ($\mathrm{GSS_E VT_E}$) for happy and aggressive. Dotted lines represent the mean of the distributions, and whiskers are set to 5th and 95th percentiles.*

4. Results and discussion

In this section, the results of both the objective and subjective experiments are presented and discussed in detail.

4.1. Objective results

Table 1 lists the mean values of the spectral distances computed from the analysed GSS and VT configurations for happy and aggressive [a] vowels. It is to note that the differences between all the configurations are statistically significant according to the Wilcoxon signed-rank test (with $p < 0.01$).

Regarding the comparison of $\mathrm{GSS_N}$ and $\mathrm{VT_N}$ with respect to $\mathrm{GSS_E}$ and $\mathrm{VT_E}$, it can be observed that GSS differences are higher for happy than for aggressive, while the opposite is happening in the case of the VT. Looking at the spectral Kullback-Leibler distances between the synthesised vowels (see the bottom of the Table 1), it can be observed that the GSS contribu-

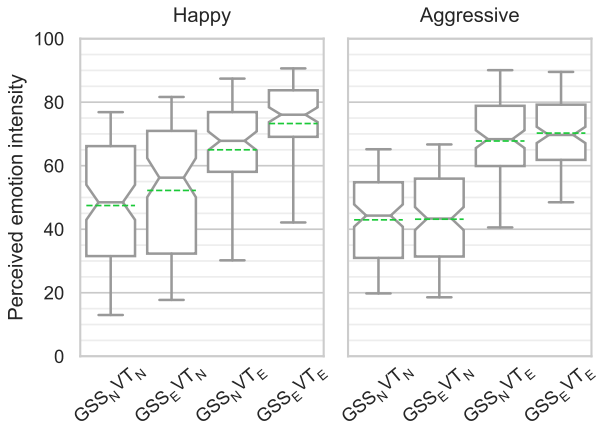


Figure 3: *Results of the MUSHRA perceptual test. Boxplots depict the perceived emotion intensity scores reported by the participants. The dotted lines represent the mean of the distributions, and whiskers are set to 5th and 95th percentiles.*

tion is more relevant in happy than in aggressive vowels. Thus, compared to the baseline, the incorporation of GSS_E is able to reduce the spectral distances to the target by 48% and 21%, respectively. The VT has a greater contribution than GSS, reducing the spectral distances by 80% for happy and by 93% for aggressive vowels.

The distributions of the computed Kullback-Leibler distances are depicted in Figure 2. It can be observed that both GSS and VT have a relevant contribution to the generation of happy and aggressive vowels, which is statistically significant according to the Wilcoxon test results. However, it is worth mentioning that VT dominates over GSS for both emotions, specially for aggressive vowels.

4.2. Perceptual evaluation results

Figure 3 depicts the results obtained from the MUSHRA test. According to the computed Wilcoxon signed-rank test, the differences between the four configurations are statistically significant with $p < 0.01$, except between $GSS_N VT_N$ and $GSS_E VT_N$ in aggressive utterances. The baseline configuration (with expressive prosody, GSS_N and VT_N) obtains the lowest perceived emotional intensity (mean score of 47 for happy and 43 for aggressive). For happy utterances, GSS_E and VT_E do significantly contribute to increase the perceived emotional intensity by 10.6% and 38.3%, respectively (from 47 to 52 and 65 points in the MUSHRA scale). When both are considered the increase is of 55.3%, thus reaching a mean score of 73 points. Regarding aggressive utterances, while GSS_E does not increase the perceived emotional intensity, VT_E leads to an increase of 58.1% (the mean score increases from 43 to 68). When both are incorporated the increase is of 62.8%, achieving a MUSHRA mean score of 70.

4.3. Discussion

Several issues can be discussed from the results obtained through the conducted objective and subjective evaluations. On the one hand, when GSS_E is applied, the distance to the target is significantly reduced for happy vowels, increasing the perceived emotion intensity for happy utterances. GSS contribution for aggressive is also objectively significant, but it is not perceived as such in the MUSHRA test unless VT_E is also transplanted. This result could be explained by analysing the differences between GSS_N and GSS_E in Table 1, which are more subtle in aggressive than in happy vowels. On the other hand, VT contribution is significantly relevant for both happy and aggressive, as observed in both the objective and perceptual analyses.

Although our work has been somehow inspired by that presented in [15], there are some important differences. Since our aim was to study the relevance of GSS and VT in resembling the target emotion, the effect of prosody has been *neutralised* as done in the second experiment of [15]. Nevertheless, in contrast to that work, GSS and VT contributions have been evaluated with a prosodic pattern coherent with the target emotion instead of using the neutral one, thereby avoiding the undesired neutralisation of the conveyed emotion as observed in the MUSHRA results. Moreover, using short utterances has not only allowed us to study vowels in their phonetic context, but also to ask evaluators about the perceived emotional intensity, instead of only evaluating isolated vowels in the arousal-valence space.

Finally, although the contribution of GSS and VT have been analysed through both objective and perceptual relative comparisons, these preliminary analyses should be completed in order to generalise the obtained results. In future works, we plan to consider more vowels and expressive styles, such as those with lax phonation, as well as other speakers covering different genders and ages.

5. Conclusions

In this work, the contribution of the GSS and VT to the generation of happy and aggressive emotional vowels has been studied on vowels [a] from a Spanish database composed of paired utterances by means of GlottDNN-based analysis and resynthesis. The objective and subjective evaluations with respect to the baseline reference (with expressive prosody, GSS_N and VT_N) show that both GSS and VT have a statistically significant contribution to convey the tense voice target emotions. Specifically, VT prevails over GSS specially for aggressive, where GSS perceptual contribution is statistically significant only when VT_E is also transplanted. Finally, it is to note that the best results are achieved when both GSS_E and VT_E are applied. When they are compared to the baseline the perceived emotional intensity is increased by 55.3% for happy and 62.8% for aggressive utterances, respectively. Properly modelling of both GSS and VT seems therefore instrumental for the upcoming 3D numerical generation of happy and aggressive vowels.

To that effect, future work will be focused on developing further analyses to extend the results obtained on vowel [a], by considering more phonemes and other expressive speaking styles and phonation types. Moreover, we envision the integration of the results within a 3D-based numerical synthesis workflow by introducing the observed relevant subtle changes in the glottal flow waveform together with the proper variations of the 3D vocal tract geometry in order to generate the desired expressive speaking style.

6. Acknowledgements

The research that has led to the results reported in this work has been funded by the SUR/DEC from the Government of Catalonia and the Ramon Llull University (ref. 2020-URL-Proj-056). The authors also would like to thank the participants on the perceptual test for their collaboration in this work.

7. References

[1] P. Taylor, *Text-to-Speech Synthesis*. Cambridge, UK: Cambridge University Press, 2009.

[2] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 537–554, 1996.

[3] M. Arnela, S. Dabbaghchian, R. Blandin, O. Guasch, O. Engwall, A. Van Hirtum, and X. Pelorson, "Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds," *The Journal of the Acoustical Society of America*, vol. 140, no. 3, pp. 1707–1718, 2016.

[4] M. Arnela, S. Dabbaghchian, O. Guasch, and O. Engwall, "Mri-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 2173–2182, 2019.

[5] M. Arnela and O. Guasch, "Finite element simulation of /asa/ in a three-dimensional vocal tract using a simplified aeroacoustic source model," in *International Congress on Acoustics (ICA)*, Aachen, Germany, sep 2019, pp. 1802–1809.

[6] R. Blandin, M. Arnela, R. Laboissière, X. Pelorson, O. Guasch, A. V. Hirtum, and X. Laval, "Effects of higher order propagation modes in vocal tract like geometries," *The Journal of the Acoustical Society of America*, vol. 137, no. 2, pp. 832–8, 2015.

[7] A. Pont, O. Guasch, and M. Arnela, "Finite element generation of sibilants /s/ and /z/ using random distributions of kirchhoff vortices," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 36, no. 2, p. e3302, 2020.

[8] M. Freixes, M. Arnela, J. C. Socoró, F. Alías, and O. Guasch, "Glottal Source Contribution to Higher Order Modes in the Finite Element Synthesis of Vowels," *Applied Sciences*, vol. 9, no. 21, p. 4535, 2019.

[9] P. Birkholz, L. Martin, K. Willmes, B. J. Kröger, and C. Neuschaefer-Rube, "The contribution of phonation type to the perception of vocal emotions in German: An articulatory synthesis study," *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1503–1512, 2015.

[10] F. Burkhardt, "Rule-Based Voice Quality Variation with Formant Synthesis," in *Proc. INTERSPEECH-2009*, Brighton, UK, 2009, pp. 2659–2662.

[11] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)*, vol. 26, no. 4, pp. 1–13, 1985.

[12] I. Yanushevskaya, C. Gobl, and C. A. Ní, "Cross-language differences in how voice quality and f_0 contours map to affect," *The Journal of the Acoustical Society of America*, vol. 144, no. 5, p. 2730, 2018.

[13] A. Murphy, I. Yanushevskaya, A. N. Chasaide, and C. Gobl, "Rd as a Control Parameter to Explore Affective Correlates of the Tense-Lax Continuum," in *Proc. InterSpeech*, Stockholm, Sweden, Aug. 2017, pp. 3916–3920.

[14] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)*, vol. 36, no. 2-3, pp. 119–156, 1995.

[15] Y. Li, J. Li, and M. Akagi, "Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space." *The Journal of the Acoustical Society of America*, vol. 144, no. 2, p. 908, 2018.

[16] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech and Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[17] O. Perrotin and I. McLoughlin, "GFM-Voc: A Real-Time Voice Quality Modification System," in *Proc. InterSpeech*, Graz, Austria, Sep. 2019, pp. 3685–3686.

[18] J. Lorenzo-Trueba, R. Barra-Chicote, T. Raitio, N. Obin, P. Alku, J. Yamagishi, and J. M. Montero, "Towards Glottal Source Controllability in Expressive Speech Synthesis," in *Proc. InterSpeech*, Portland, OR, USA, Sep. 2012, pp. 1620–1623.

[19] M. Airaksinen, L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "A comparison between straight, glottal, and sinusoidal vocoding in statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1658–1670, Sep. 2018.

[20] S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, "Vocal effort based speaking style conversion using vocoder features and parallel learning," *IEEE Access*, vol. 7, pp. 17 230–17 246, 2019.

[21] M. Freixes, M. Arnela, F. Alías, and J. C. Socoró, "GlottDNN-based spectral tilt analysis of tense voice emotional styles for the expressive 3D numerical synthesis of vowel [a]," in *Proc. 10th ISCA Speech Synthesis Workshop (SSW)*, Vienna, Austria, Sep. 2019, pp. 132–136.

[22] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[23] I. Iriondo, S. Planet, J.-C. Socoró, E. Martínez, F. Alías, and C. Monzo, "Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification," *Speech Communication*, vol. 51, no. 9, pp. 744–758, 2009.

[24] L. Rabiner, *Fundamentals of speech recognition*. PTR Prentice Hall, 1993.

[25] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 39–51, 2001.

[26] R. ITU, "ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems," *International Telecommunication Union*, 2003.

[27] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[28] N. Jillings, B. De Man, D. Moffat, and J. D. Reiss, "Web audio evaluation tool: A browser-based listening test environment," in *12th International Conference in Sound and Music Computing (SMC 2015)*. Maynooth, Ireland: SMC network, Jul. 2015, pp. 147–152.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

The age effects on EP vowel production: an ultrasound pilot study

Luciana Albuquerque[1,2,3,4], *Ana Rita Valente*[1,3], *Fábio Barros*[1,3], *António Teixeira*[1,3], *Samuel Silva*[1,3], *Paula Martins*[1,5,6], *Catarina Oliveira*[1,5]

[1]Institute of Electronics and Informatics Engineering of Aveiro, Aveiro, Portugal
[2]Center for Health Technology and Services Research, University of Aveiro, Portugal
[3]Dep. Electronics, Telecommunications and Informatics, University of Aveiro, Portugal
[4]Department of Education and Psychology, University of Aveiro, Aveiro, Portugal
[5]School of Health Sciences, University of Aveiro, Portugal
[6]Institute of Biomedicine, University of Aveiro, Portugal
{lucianapereira, rita.valente, fabiodaniel, ajst, sss, pmartins, coliveira}@ua.pt

Abstract

For aging speech, there is a limited knowledge regarding the articulatory adjustments underlying the acoustic findings observed in previous studies. In this context, ultrasound imaging is a technology that can be safely used to study static and dynamic features of the articulators allowing comparisons of physiological differences between older and young adults during speech production. In order to investigate the age-related articulatory differences in European Portuguese vowels, the present study analyzes the tongue contours of the 9 European Portuguese oral vowels in isolated context and in a disyllabic sequence. From the tongue contours segmented from the Ultrasound images, several parameters were extracted (e.g., tongue height, tongue advancement) to allow comparisons between speakers of different age groups. For this study, while the analysis of data for more speakers is ongoing, we considered a set of four European Portuguese native female speakers of two different age groups and addressed the study of the oral vowels articulatory space. The preliminary results suggest that the vowel articulatory space, namely for disyllabic sequences, tend to be smaller in the older females.

Index Terms: 2D-Ultrasound tongue imaging, European Portuguese vowels, Speech Production, Aging speech

1. Introduction

The aging process causes specific alterations in the speech organs (e.g., stiffening of thorax, decreased lung capacity, weakening of respiratory muscles and atrophy of facial, mastication and pharyngeal muscles) [1, 2, 3], and all these changes are expected to play an important role in speech production. Over the years, although age-related variations on the acoustic properties of speech have been extensively investigated [2, 3, 4], its underlying articulatory details have not been well understood. On what concerns age-related changes on vowel formant frequencies (mostly F1 and F2), the results across studies are highly inconsistent [2, 3, 5, 6]. Previous acoustic results for European Portuguese (EP) were also not conclusive, as there are vowels that presented a different pattern of formant frequencies variation with age and gender [7, 8, 9]. [8, 10] observed that vowel formants tend to decrease mainly in females and to centralize in older males with aging, and these changes might be related to specific articulatory adjustments of the older speakers during speech.

Unlike acoustic studies, in which age-related speech variations have been widely studied since the 1960s [3], on what concerns articulatory studies there are only a few studies about effects of aging on tongue position and strength during speech production [11, 12, 13, 14]. Ultrasound (US) tongue imaging synchronized with audio can be used to investigate the physiological differences between old and young adult speech. For this reason, the main objective of the ongoing study is to investigate the age-related articulatory differences in EP vowels with US imaging. Additionally, since there is a paucity of literature on EP oral vowel production, the available data were collected mainly for acoustic studies [7, 8, 15, 16] or in articulatory studies of nasal vowels [17, 18, 19], this study also provides valuable insights to a first articulatory description of all EP oral vowels with US.

Due to the limited knowledge of the articulatory basis of previously acoustic findings concerning aging speech, the aim of this pilot study is to analyze and compare the US tongue contours of the 9 EP oral vowels in isolated context and in a disyllabic sequence produced by four female speakers between 19 and 66 years, to infer preliminary results of the age effect on EP vowel production, which is essential for the development of automatic speech recognition (ASR) systems suitable for older speech, and for clinical assessment and treatment of speech disorders.

2. Background and Related Work

Speech production has been studied essentially through speech sound acoustics (for EP: [7, 8, 15, 16]). Although frequencies are an indirect measure of articulatory movements, direct measures can be obtained over different articulatory techniques, such as electromagnetic articulography (EMA), real-time magnetic resonance (RTMRI) and US imaging (for EP: [17, 18, 19]). US imaging presents several advantages in comparison with the other referred techniques: it is a non-invasive, safe, portable and fast technology that is commonly used to demonstrate the midsagittal surface contour of the tongue [20], and it can contribute with important information for different areas in speech research [21, 22].

Despite US being a more affordable alternative for several contexts, enabling the acquisition of larger datasets, it demands adequate computational approaches for processing and analysis. US artifacts, corrupting noise, and the presence of spurious edges are also challenges for the processing of US images [22, 23]. Furthermore, a challenge in measuring US images is the lack of a physiological reference [23]. Even though the tongue contours are visible, there are no hard structure references (i.e., US does not image internal articulators other than the tongue), making it difficult to determine an exact position for the tongue in the vocal tract [24]. The head and transducer holders help overcome these problems [23], but cannot be guaranteed to be re-fitted to

10.21437/IberSPEECH.2021-52

the same location on a speakers' head in different trials [25]. For that, re-orienting images to a common co-ordinate system with a bite plane allows for some degree of normalization and it is more tolerant to error in the placing of the probe outwith the midsagittal plane [25].

The lack of reference points and the anatomical differences also introduce a difficulty in comparing speech data across speakers, that is, comparing lingual articulation across age groups raises problems of normalization [24, 26]. This issue has been also recognized in the acoustic studies of vowel formants; and indeed, the length of the vocal tract can be considered a greater influence in the articulatory field, leading to the need of normalization procedures [26]. External references or the image of internal articulators have been proposed as solutions [24, 23]. However, there is no commonly accepted method for comparing tongue shapes among speakers [27].

Several research studies developed since 2010 have been used US tongue images to investigate, on different language systems, the articulatory correlates of vowels production [28, 29]. Concerning the articulatory parameters, tongue contour and also height and advancement of the highest point of the tongue had been successfully used in previous US studies with vowels.

To the best of our knowledge, the articulatory studies concerning vowels' properties across lifespan are scarce and the majority focuses on coarticulation issues. An articulatory study with 3D electromagnetic articulography suggested that especially the tongue body was affected by age, and the movements for the vowels were slower in the older speakers compared to the younger ones [11]. The results of an US study of antecipatory velar-vowel coarticulation and speech stability in speakers who do and do not stutter across lifetime [13] indicated an age effect, with progressive less coarticulation and increase speech stability with aging. Regarding the performance on motor tasks with increasing age, a generalized slowness, decrease coordination, a diminished performance level and on precise motor control were observed [13, 11], which could affect vowel production.

3. Method

This cross-sectional study was approved by the Ethics Committee of Escola Superior de Enfermagem de Coimbra, Portugal (approval number 639/12-2019), and all participants agreed and signed the consent form before participating in the study.

3.1. Speakers and Corpus

Ultrasound data were collected from a convenience sample of four EP native female speakers, two young (S1 - 19 years; S2 - 31 years) and two old females (S3 - 63 years; S4 - 66 years), to test ages with more expected distinctive characteristics. Data collection took place in a pandemic year, which conditioned the acquisition of a larger number of participants, namely older people. All of them were in good health and with no reported history of neurological disorders or diseases, or any speech, language or hearing difficulties. Considering the speakers' anatomical characteristics, S1 speaker is 174 cm tall and weighs 95 kg; S2 is 171 cm tall and weighs 56 kg; S3 is 152 cm tall and weighs 59 kg; and S4 is 150 cm tall and weighs 52 kg.

The corpus consisted of all EP oral vowels [i], [e], [ɛ], [a], [o], [ɔ], [u], [ɨ] and [ɐ] in pseudoword context and in isolated context. The pseudoword list contained ˈpV.Cv sequences (started with the labial voiceless stop consonant [p]), where C was balanced for the place of articulation using the voiceless stop consonants [p], [t] and [k], and V was all EP oral vowels in stressed position. The last vowel (i.e., v) corresponds only to the vowels [u], [ɨ] and [ɐ]. The stimuli were embedded in a carrier sentence "Em pVCv temos V" (*In pVCv we have V.*). For each vowel, three different pseudowords were selected. Each carrier sentence was repeated 3 times. Thus, each speaker produced 81 individual utterances, i.e., 9 repetitions of each vowel, per context.

3.2. Data Acquisition

The participants were asked to seat, facing a computer screen displaying prompts, and to wear a stabilization helmet [30], in order to ensure that neither the speaker's head nor the transducer moved during the experiment.

Synchronous acquisition of US images and speech sounds through Articulate Assistant Advanced software (AAA) [31] took place in quiet rooms using an endocavitary probe (65EC10EA) with 90º field of view positioned under the participants' chin. US images were collected with an US machine Mindray DP6900 at a frame rate of 60 Hz and the depth setting was 97 mm. Audio was collected with a Philips SBC ME400 microphone connected to an external sound system (UA-25 EX USB). The recorded data was collected as video and audio synchronized with a SyncBrightUp unit [32].

Instructions were provided prior to recording to ensure familiarity with the speech materials. The speech material was presented in three randomized blocks (i.e., front ([i], [e], [ɛ]), central ([ɨ], [ɐ], [a]) and back ([u], [o], [ɔ]) vowels). Each block began and finished with the production of the sequence /tatatata/ to assess sound and image synchronization. Also, at the start of each block a recording of the bite plane was obtained in order to image the occlusal plane, which is a reliable method for the definition of horizontal and vertical orientations in the vocal tract [25, 33]. That is, the speaker was asked to bite and press their tongue against a flat plastic plate, which results in their tongue bulging upward at the back edge of the bite plate [25, 33].

Taking into consideration different anatomical characteristics of each speaker and that the optimal probe orientation is vowel dependent, its orientation was adjusted, along the sessions, for each block of vowels, to enable the best possible imaging of the tongue. The bite plane sequences were then used as a common referential, for each speaker.

3.3. Data processing and statistical analyses

Data processing – The acoustic files were exported from AAA software in WAV format for automatic segmentation at word and phoneme level using WebMAUS [34], and then imported into Praat software [35], to manually check the accuracy of the vowel boundaries.

To reduce the impact of the noisy nature of the US images, a pre-processing was applied and consists in: 1) cropping the US image to remove irrelevant information and to select a region of interest (ROI); 2) applying a phase symmetry filter to the ROI to enhance the outline corresponding of the tongue surface [36, 37]; 3) applying a radial sweep approach with 5º of angular distance [29]; 4) collecting of all pixel intensities; 5) extraction of the highest intensity point for each radial sweep. More details in [38, 39]. Given the challenging nature of the images, and to ensure the reliability of the data, the tongue segmentation of all frames of the vowel occurrences was revised by three annotators with experience in speech production analysis (see Fig. 1).

For each block of vowels, tongue contours were rotated to the speaker's bite plane obtained in the corresponding block, so that the image of the occlusal plane was observed to be parallel to the upper and lower edges of the video pane [25, 26, 33].

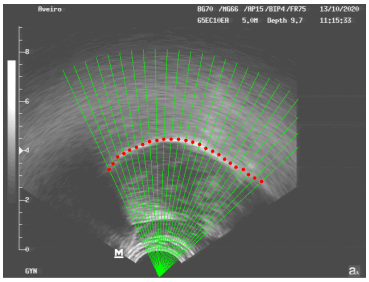


Figure 1: *Illustrative US image with radial sweeps and the extracted points on the tongue contour for vowel [a]*

Data were exported after rotation and the origin of the coordinates system corresponds to the back of the bite plate, 4 cm from the upper incisors along the occlusal plane (see the bite plane traces in Fig. 2).

For the present study, only the tongue contour of the temporal midpoint of the vowels, which consistently contained an articulatorily steady part of the vowel [33], were exported in cartesian coordinates. Ideally, the total number of vowels analyzed per speaker was 81, which corresponds to 81 frames. However some frames had to be discarded from the analysis due to poor US image quality for some speakers (namely for S3), vowels (mostly back vowels), and/or context (mostly in isolated context). Furthermore, due to the fact that some tongue contours were not totally visible, some tokens were segmented incomplete and not being able to find important descriptors for vowel analysis, such as the highest point of the tongue body.

Articulatory Measures – Each vowel tongue contours per speaker and context were obtained as the median value of all vowel occurrences. The highest point of the tongue's contour was extracted and represents the highest point of the tongue body (i.e., tongue height, TH). The x-coordinate reflects the front back position of the tongue in the y coordinate (i.e., tongue advancement, TA). Pixel to cm conversion was made considering 1 cm corresponds to 44 pixels in the image. The TH and TA for each vowel, per context and speaker, was obtained based on the median value of all tokens at the temporal midpoint, which reduces the effect of the flanking consonants and the effect of measurement errors.

4. Results

In this section a summary of the main findings for the analysis of the vowel tongue configuration on EP vowels are presented. Also some considerations about inter and intra-speaker differences are reported. Due to space limitations, contours and vowel articulatory spaces of all speakers are not given here, and only relevant differences are presented. Note that, for speaker S3, none completed contour of the back vowels [o] and [ɔ] in isolated context was obtained, which affect the TH and TA measures presented for this speaker.

Tongue contours – Concerning the tongue contours analysis, results for one speaker, both contexts, are presented in Fig. 2. The differences in tongue contours appear to be clearly vowel dependent in isolated vowels. For vowels in pVCv sequences the contours for different vowels tend to be more similar.

Articulatory Measures – Considering the tongue contours, the highest point of the tongue body was determined and analyzed. Fig. 3 summarizes the TH and TA values obtained for each speaker by vowel in pVCv sequence, and in isolated context. In general, vowels in pVCv sequence presented higher variability of TA and TH values than isolated vowels due to the influence of the consonantal context [40, 41]. Regardless of the vowel

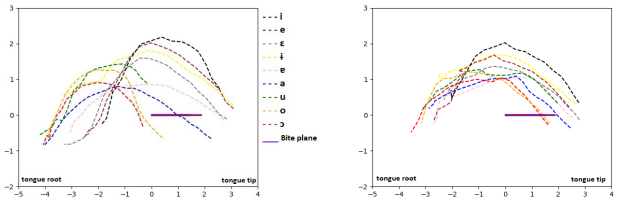


Figure 2: *Tongue contours for all vowels in isolated (left side) and in pVCv sequences (right side) for speaker S4 (scale in cm)*

context, for each speaker, the higher TH was observed for vowel [i], and the lowest TH was obtained, mostly, for the vowel [a]. The lowest TA was obtained, mostly, for the vowel [o]. Concerning the vowel that presented the highest TA, data showed more variability, with some speakers present vowel [ɐ] with higher TA than expected, mostly in isolated context. Concerning the highest TA, as data showed greater variability, no vowel can be pointed.

Articulatory vowel space – For intra-speaker comparisons, Fig 4 represents the articulatory space defined by TA and TH of the cardinal EP oral vowels ([a], [i] and [u]) for each speaker, in isolated and in pVCv context. Each vowel is represented by the median TH and TA. Comparing the TH and TA of the tongue contours of these female speakers, it can be observed that articulatory space tends to be smaller when vowels occur in pVCv sequences comparing with isolated vowels. The data also tends to indicate that the vowel articulatory space differences observed between both contexts is higher for the old females. That is, the articulatory vowel space area, namely for pVCv sequences, tend to be smaller in the older females.

Articulation variability – Plots in Fig. 4 only provide the average, but variability of productions is also very important to analyze. Fig. 5 represents individual productions and information regarding dispersion based on ellipses. Fig. 5 represents the TH and TA of the total number of occurrences of the cardinal vowels in both contexts for two speakers (one young and one old). In those graphs it can observed that the dispersion of isolated vowels is lower than in pVCv sequences, mainly for vowel [a]. Also, in both contexts, for these two speakers in can be observed that vowel [a] presents the highest variability.

5. Conclusions

This paper presents the initial results of the automatic extraction contours and the determination of the highest position of the tongue body for young and old Portuguese females. The method revealed being able to detect articulatory measures for young and old speakers, making possible the production of the first vowel articulatory space representation for EP. Regarding limitations, this study presents a reduced sample that, due non-normalization procedures, hinder inter-speaker comparisons. Also, the noisy nature of the images make the segmentation demanding and could difficult the accurate determination of articulatory measures for some vowels. These issue could be reflected in the greater variability observed for certain vowels.

The vowel articulatory space reduction observed for isolated vowels in comparison with the vowels in pVCv sequences might be related with the tendency to hyperarticulate isolated vowels. So, this vowel articulatory space reduction was similar between vowels in clear speech versus in conversational speech [40], or in long vowels versus short vowels [42], for other languages. Also the tendency to hyperarticulate isolated vowels might be in the origin of the more distinct tongue contours for these vowels.

Future work – Ensure inter-speaker comparison through the aplication of normalization procedures; improvement of

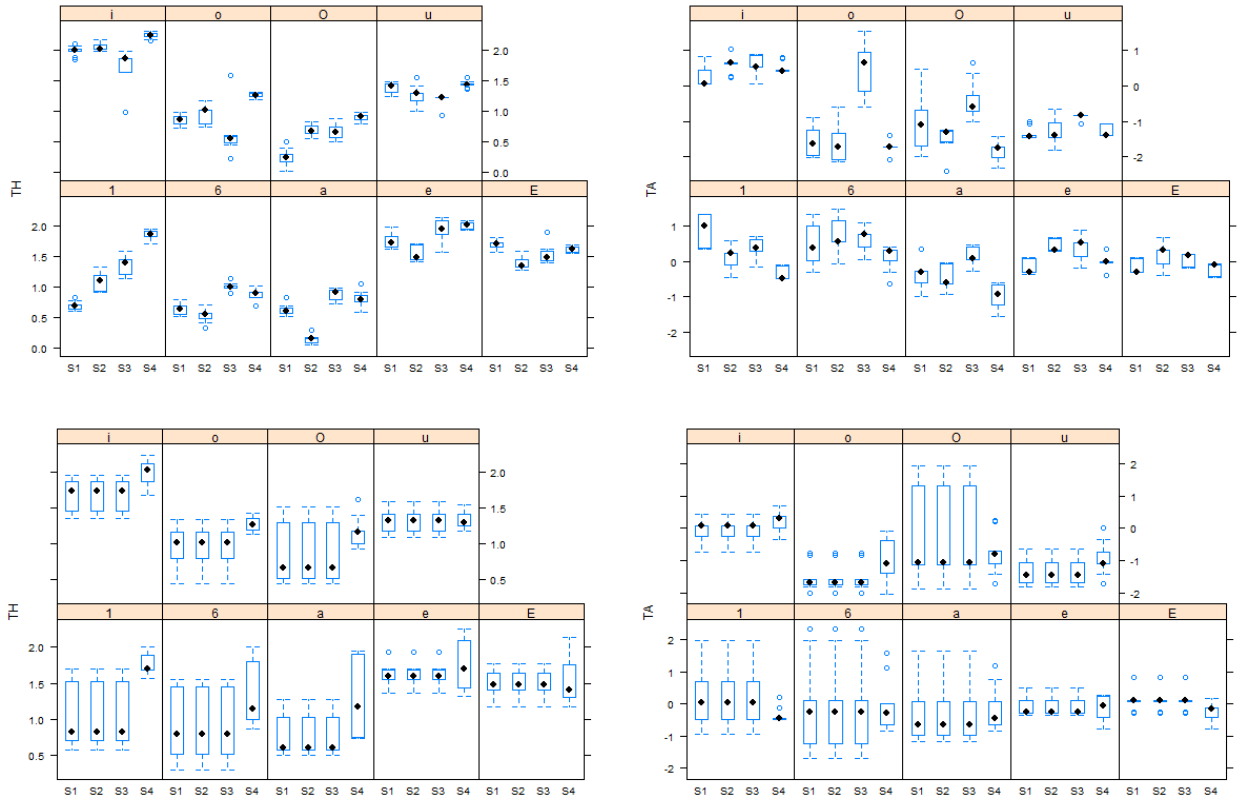


Figure 3: *TH e TA by vowel type and speaker (SAMPA notation). Top: isolated vowels; Bottom: vowels in pVCv sequence (scale in cm)*

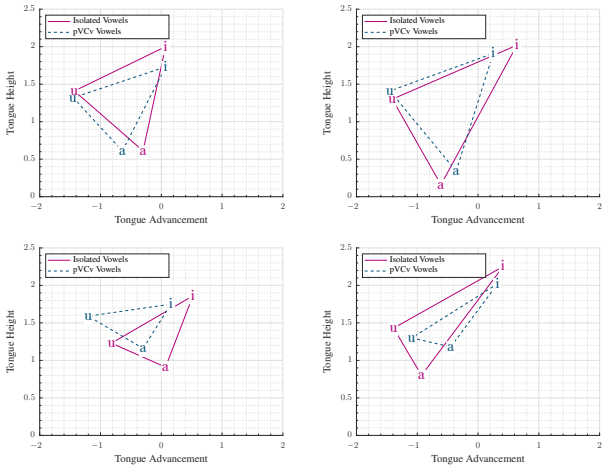


Figure 4: *Articulatory vowel space of the EP cardinal vowels in isolated (pink solid lines) and in pVCv context (blue dashed lines). Top: young females (S1 and S2); bottom: old females (S3 and S4)*

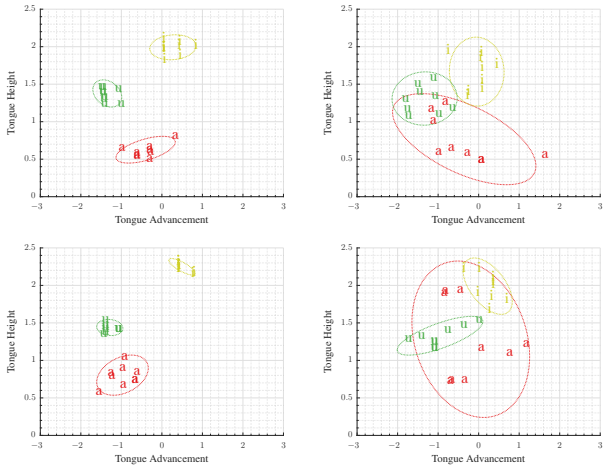


Figure 5: *Vowel articulatory cluster size of the EP cardinal vowels of two speakers in isolated vowels (left side) and in pVCv sequences (right side). Top: young female (S1); Bottom: old female (S4)*

the determination of the highest point of the tongue; explore SSANOVA to study vowel contours; investigate articulatory movement and velocity; a large study of the age effect on vowel articulation measures.

6. Acknowledgements

This research was financially supported by the projects VoxSenes (POCI-01-0145-FEDER-03082) and MEMNON (POCI-01-0145-FEDER-028976) – COMPETE2020 under POCI and FEDER, and by national funds (OE), through FCT/MCTES –, SOCA – Smart Open Campus CENTRO-01-0145-FEDER-000010 (Portugal 2020 under POCI and FEDER) and by IEETA Research Unit funding (UIDB/00127/2020). First author was funded by FCT grant SFRH/BD/115381/2016.

7. References

[1] P. Massimo and P. Elisa, "Age and Rhtymic Variations: A study on Italian," in *INTERSPEECH*, Singapore, 2014, pp. 1234–1237.

[2] S. E. Linville, *Vocal aging*. Australia, San Diego: Singular Thomson Learning, 2001.

[3] S. Schötz, *Perception, analysis and synthesis of speaker age*. Lund University: Linguistics and Phonetics, 2006, vol. 47.

[4] R. Vipperla, S. Renals, and J. Frankel, "Ageing voices: The effect of changes in voice parameters on ASR performance," *EURASIP J. Aud. Speech Music Process*, pp. 1–10, 2010.

[5] M. P. Rastatter, R. A. McGuire, J. Kalinowski, and A. Stuart, "Formant frequency characteristics of elderly speakers in contextual speech," *Folia Phoniatrica et Logopaedica*, vol. 49, no. 1, pp. 1–8, 1997.

[6] J. T. Eichhorn, R. D. Kent, D. Austin, and H. K. Vorperian, "Effects of Aging on Vocal Fundamental Frequency and Vowel Formants in Men and Women," *Journal of Voice*, vol. 32, no. 5, pp. 644.e1–644.e9, 2018.

[7] L. Albuquerque, C. Oliveira, A. Teixeira, P. Sa-Couto, J. Freitas, and M. S. M. Dias, "Impact of age in the production of European Portuguese vowels," in *INTERSPEECH*, Singapore, 2014, pp. 940–944.

[8] L. Albuquerque, C. Oliveira, A. Teixeira, P. Sa-Couto, and D. Figueiredo, "A comprehensive analysis of age and gender effects in European Portuguese oral vowels," *Journal of Voice*, no. In press, dec 2020.

[9] T. Pellegrini, A. Hämäläinen, P. B. de Mareüil, M. Tjalve, I. Trancoso, S. Candeias, M. S. Dias, and D. Braga, "A corpus-based study of elderly and young speakers of European Portuguese: acoustic correlates and their impact on speech recognition performance," in *INTERSPEECH*, Lyon, 2013, pp. 852–856.

[10] L. Albuquerque, C. Oliveira, A. Teixeira, P. Sa-Couto, and D. Figueiredo, "Age-related changes in European Portuguese vowel acoustics," in *INTERSPEECH*, Graz, 2019, pp. 3965–3969.

[11] A. Hermes, J. Mertens, and D. Mücke, "Age-related Effects on Sensorimotor Control of Speech Production," in *INTERSPEECH*, Hyderabad, 2018, pp. 1526–1530.

[12] P. De Decker and S. Mackenzie, "Tracking the phonological status of /l/ in Newfoundland English: Experiments in articulation and acoustics," *J. Acoust. Soc. Am.*, vol. 142, no. 1, pp. 350–362, 2017.

[13] A. J. Belmont, "Anticipatory Coarticulation and Stability of Speech in Typically Fluent Speakers and People Who Stutter Across the Lifespan: An Ultrasound Study," Master of Science, University of South Florida, 2015.

[14] A. T. Neel and P. M. Palmer, "Is Tongue Strength an Important Influence on Rate of Articulation in Diadochokinetic and Reading Tasks?" *JSLHR*, vol. 55, pp. 235–246, 2012.

[15] P. Escudero, P. Boersma, A. S. Rauber, and R. A. H. Bion, "A cross-dialect acoustic description of vowels: Brazilian and European Portuguese," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1379–1393, 2009.

[16] C. Oliveira, M. M. Cunha, S. Silva, A. Teixeira, and P. Sa-Couto, "Acoustic analysis of European Portuguese oral vowels produced by children," in *IberSPEECH*, vol. 328, Madrid, 2012, pp. 129–138.

[17] C. Oliveira, P. Martins, S. Silva, and A. Teixeira, "An MRI study of the oral articulation of European Portuguese nasal vowels," in *13th INTERSPEECH*, Portland, OR, USA, 2012, pp. 2690–2693.

[18] C. Cunha, S. Silva, A. Teixeira, C. Oliveira, P. Martins, A. Joseph, and J. Frahm, "On the Role of Oral Configurations in European Portuguese Nasal Vowels," in *INTERSPEECH*, Graz, Austria, 2019, pp. 3332–3336.

[19] C. Oliveira, P. Martins, and A. Teixeira, "Speech rate effects on European Portuguese nasal vowels." in *INTERSPEECH*, Brighton, 2009, pp. 480–483.

[20] L. Lancia, P. Rausch, and J. S. Morris, "Automatic quantitative analysis of ultrasound tongue contours via wavelet-based functional mixed models," *J. Acoust. Soc. Am.*, vol. 137, no. 2, pp. EL178–EL183, 2015.

[21] M. H. Mozaffari, S. Wen, N. Wang, and W. Lee, "Real-time automatic tongue contour tracking in ultrasound video for guided pronunciation training," in *14th VISIGRAPP*, vol. 1, 2019, pp. 302–309.

[22] Y. S. Akgul, C. Stone, and K. Maureen, "Automatic extraction and tracking of contours," *TRANSACTIONS ON MEDICAL IMAGING*, vol. 18, no. 10, pp. 1035–1045, 1999.

[23] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clin Linguist Phon*, vol. 19, no. 6-7, pp. 455–501, 2005.

[24] N. Zharkova, N. Hewlett, and W. J. Hardcastle, "Coarticulation as an indicator of speech motor control development in children: An ultrasound study," *Motor Control*, vol. 15, no. 1, pp. 118–140, 2011.

[25] J. M. Scobbie, E. Lawson, S. Cowen, J. Cleland, and A. A. Wrench, "A common co-ordinate system for mid-sagittal articulatory measurement," in *QMU CASL Working Papers WP-20*, Edinburgh, 2011.

[26] P. Strycharczuk and J. M. Scobbie, "Fronting of Southern British English high-back vowels in articulation and acoustics," *J. Acoust. Soc. Am.*, vol. 142, no. 1, pp. 322–331, 2017.

[27] G. Barbier, P. Perrier, L. Ménard, Y. Payan, M. Tiede, and J. Perkell, "Speech planning in 4-year-old children versus adults: Acoustic and articulatory analyses," in *INTERSPEECH*, 2015.

[28] K. Comivi Alowonou, J. Wei, W. Lu, Z. Liu, K. Honda, and J. Dang, "Acoustic and Articulatory Study of Ewe Vowels: A Comparative Study of Male and Female," in *INTERSPEECH*. Graz, Austria: ISCA, 2019, pp. 1776–1780.

[29] L. Ménard, C. Toupin, S. R. Baum, S. Drouin, J. Aubin, and M. Tiede, "Acoustic and articulatory analysis of French vowels produced by congenitally blind adults and sighted adults," *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 2975–2987, 2013.

[30] Articulate Instruments Ltd., "Ultrasound stabilisation headset users manual," Edinburgh, UK, 2008.

[31] Articulate Assistant Ltd., "Articulate Assistant Advanced ultrasound module user manual," 2014.

[32] Articulate Instruments Ltd., "SyncBrightUp users manual," Edinburgh, UK, 2010.

[33] M. Dokovova, M. Sabev, J. M. Scobbie, R. Lickley, and S. Cowen, "Bulgarian vowel reduction in unstressed position: an ultrasound and acoustic investigation," in *19th ICPhS*, 2019, pp. 2720–2724.

[34] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech and Language*, vol. 45, pp. 326–347, 2017.

[35] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," University of Amsterdam, 2012.

[36] P. Kovesi *et al.*, "Symmetry and asymmetry from local phase," in *Tenth Australian joint conference on artificial intelligence*, vol. 190. Citeseer, 1997, pp. 2–4.

[37] E. Karimi, L. Ménard, and C. Laporte, "Fully-automated tongue detection in ultrasound images," *Computers in Biology and Medicine*, vol. 111, no. 103335, pp. 1–13, 2019.

[38] F. Barros, A. R. Valente, L. Albuquerque, S. Silva, A. Teixeira, and C. Oliveira, "Contributions to a quantitative unsupervised processing and analysis of tongue in ultrasound images," *Lecture Notes in Computer Science*, vol. 12132 LNCS, pp. 170–181, 2020.

[39] F. Barros, S. Silva, L. Albuquerque, A. R. Valente, A. Teixeira, P. Martins, and C. Oliveira, "Towards the use of ultrasonography to study aging effects in vowel production," in *12th ISSP*, 2020.

[40] J. Y. Song, "The use of ultrasound in the study of articulatory properties of vowels in clear speech," *Clin Linguist Phon*, vol. 31, no. 5, pp. 351–374, 2017.

[41] N. Zharkova, F. E. Gibbon, and W. J. Hardcastle, "Quantifying lingual coarticulation using ultrasound imaging data collected with and without head stabilisation," *Clinical linguistics & phonetics*, vol. 29, no. 4, pp. 249–265, 2015.

[42] W.-S. Lee, "Articulatory–Acoustical Relationship in Cantonese Vowels," *Language and Linguistics*, vol. 17, no. 4, pp. 477–500, 2016.

Exploring Transformer-based Language Recognition using Phonotactic Information

David Romero[1], *Luis Fernando D'Haro*[2], *Christian Salamea*[1,2]

[1]Interaction, Robotics and Automation Research Group, Universidad Politécnica Salesiana, Calle Vieja 12-30 y Elia Liut, Cuenca, Ecuador.
[2]Speech Technology Group, Information and Telecommunication Center, Universidad Politécnica de Madrid, Ciudad Universitaria Avda. Complutense, 30, 28040, Madrid
dromeromog@gmail.com, lfdharo@die.upm.es, csalamea@ups.edu.ec

Abstract

This paper describes an encoder-only approach based on the "Transformer architecture" applied to the language recognition (LRE) task using phonotactic information. Due to the use of one global set of phonemes to recognize all languages, the proposed system needs to overcome difficulties due to the overlapping and high co-occurrences of similar phone sequences across languages. To mitigate this issue, we propose a single transformer-based encoder trained for classification, where the attention mechanism and its capability of handling large sequences of phonemes help to find discriminative sequences of phonotactic units that contribute to correctly identify the language for short, mid and long audio segments. The proposed approach provides significant improvements, outperforming phonotactic-based RNNs and Glove-based i-Vectors architectures, getting a relative improvement of 5.5% and 38.5% respectively. Our experiments were carried out using phoneme sequences obtained by the "Allosaurus phoneme recognizer" applied to the Kalaka-3 Database. This dataset is challenging since the languages to identify are mostly similar (i.e. Iberian languages, e.g. Spanish, Galician, Catalan). We provide results using the C_{avg} metric proposed for NIST evaluations.

Index Terms: language recognition, phonotactic information, self-attention mechanism.

1. Introduction

Phonotactic-based LRE attempts to recognize the language that is spoken in a speech sample using as input the sequence of phonemes obtained by a phoneme recognizer, which uses a global set of phonemes to transcribe each audio file in the corpus with independency of the languages to be recognized. This global set contributes to make the system easier, in contrast with traditional approaches like Parallel Phone Recognition Language Modeling (PPRLM) [1], but introduces difficulties due to the overlapping and high co-occurrences of similar phonemes between languages. Therefore, finding discriminative combinations of these units is one of the main goals to differentiate transcriptions generated by similar languages.

Recent approaches to phonotactic LRE have used Recurrent Neural Networks (RNN's) [2][3] to take advantage of its recurrent ability for processing sequences using past information. However, if we consider that these phonetic sequences are usually long, this type of neural network could suffer from vanishing gradient problems [4] during training, which lead to performance degradation. Other approaches [5] have used i-Vector-based frameworks that take as input phonotactic-based embeddings learned through Skip-Gram [6] and Glove [7] models. While powerful, these approaches use small context windows limiting the use of longer context information that could be useful to find discriminative combinations.

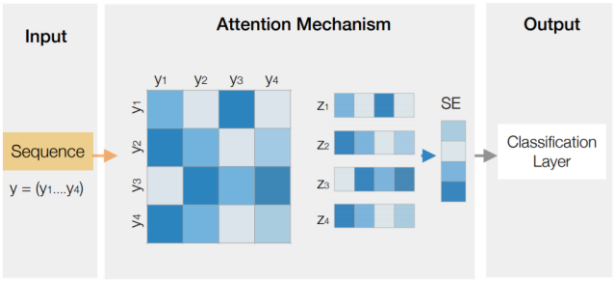


Figure 1: *High level overview of our approach. The attention mechanism is used to find discriminative combinations of phonetic units.*

In this work, we consider the use of an encoder-only approach based on the Transformer architecture [8] for the LRE task using phonotactic information. Models based on Transformers have shown their versatility and robustness [9] to perform various tasks such as seq2seq translation [8], sentiment analysis [10], or language understanding [11], commonly achieving state-of-the-art results outperforming architectures like Long Short-Term Memory (LSTM's) [12]. We attempt to take advantage of its self-attention mechanism that allows the network to use and capture contextual information from long sequences, helping to find discriminative combinations of the global phonetic units.

Figure 1 shows an overview of our approach. We use as input an embedding sequence $(y_1,..,y_L)$, where each embedding represents a n-gram phonetic unit of a given input sequence. Then, contextual vector representations of these units are learned together with the self-attention weights, allowing each unit to attend to the other units in the sequence capturing contextual information and helping to find the most relevant tokens to make each final representation $(z_1,...,z_L)$. Finally, we use average pooling to obtain the final sentence embedding (SE) which is the input to a classification layer. Our experimental results show that the proposed architecture outperforms RNN's and i-Vector based systems for modeling phonotactic information only.

10.21437/IberSPEECH.2021-53

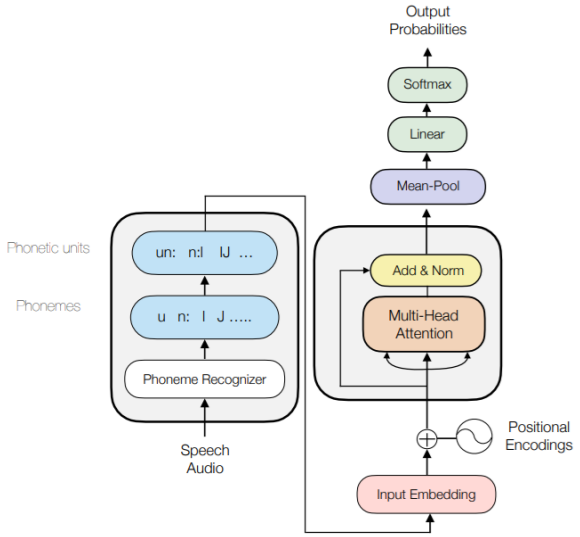


Figure 2*: An overview of the Transformer-based Encoder applied for LRE using phonotactic information.*

This paper is organized as follows: in Section 2 describes the proposed architecture. Section 3 explains the experimental setup and describes the database and metrics used in our experiments. Section 4 discusses the obtained results. And, finally, Section 5 shows the conclusions and future work.

2. Model Description

In this section we describe the proposed model in detail. Figure 2 shows the different components of the architecture with special emphasis on the simplified transformer-based encoder used in this paper. Below, we provide detailed information for each component in the figure.

2.1. Phoneme recognizer

The first step is the recognition of phoneme sequences for each audio file in the corpus, using the "Allosaurus Phoneme recognizer" [13]. This recognizer incorporates some phonology knowledge in its model through an allophone layer, which associates a universal narrow phone set with the phonemes that appear in each language, making it possible to perform universal phone recognition. This allows the model to incorporate individual recognizers for over 2000 languages and a global recognizer called "IPA" that use a vocabulary of around 230 phonemes using the inventory of all the languages supported in the recognizer, which includes the languages of the corpus that is used in this work.

For this task we use a different approach than other works based on RNN's [2,3] and i-Vectors [5], that use a phoneme recognizer developed by the Brno University of Technology (BUT) [14], which has only 3 recognizers for Hungarian, Russian and Czech with 61, 46 and 52 different phonemes respectively. These approaches use only one of these recognizers to make phoneme representations for all the languages in the corpus. In our case, we use the Allosaurus phoneme recognizer using the global vocabulary, with which we obtain better performance; on the one hand, due to the better recognition using a global set of phonemes that summarizes all the languages in the corpus rather than using a single model that incorporates information of only one of them; on the other hand, it has a larger phoneme vocabulary set than the Brno Recognizer, allowing more variability in the transcriptions (this will be explained in more detail in the next subsection) which ended showing benefits to the transformer-based encoder. We believe this variability helps the model to find unique tokens and therefore discriminative combinations that benefit the LRE task.

2.2. Phonetic units

To incorporate more context information as input to the architecture, we used n-gram phonetic units. The use of these units has shown better performances [4] than just using phonemes thanks to the incorporation of local information and diversity when learning the vector embeddings for each unit. Let $x = (x_1, x_2, \dots, x_L)$ be a sequence of recognized phonemes of length L; we form phonetic units by concatenating two or more phonemes into one new n-gram unit, generating the new sequence $x_n = (x_{1:2}, x_{2:3}, \dots, x_{L-1:L})$, the start of sentence (SOS) and end of sentence (EOS) tokens have been added at the beginning and end of this sequence that will be used as input to the transformer-based encoder. We experimented with different sizes for these phonetic units (n-gram order) which will be presented in the next section. The creation of these units is shown in the top left block of Figure 2.

2.3. Transformer-based Encoder

The original transformer [8] consists of stacked encoder and decoder layers. In this work we have used only an encoder-based approach trained for classification. In this subsection we describe the encoder layer and how it was applied for our work, please refer to Figure 2 for this discussion.

2.3.1. Input

The input of the encoder layer is a sequence of phonetic units $x = (x_{1:2}, x_{2:3}, \dots, x_{L-1:L})$, each unit in this sequence is tokenized using a "Word tokenizer". Once this tokenization is done, each token is represented by an embedding that is learned during training, this input sequence is represented as $y = (y_{1:2}, y_{2:3}, \dots, y_{L-1:L})$ where $y_i \in \mathbb{R}^{d_{model}}$.

2.3.2. Positional Encodings

Since the transformer contains no recurrence to make use of the order of the sequence, it uses "Positional encodings" that are added to the input embeddings at the bottom of the layer. The positional encodings have the same dimension d_{model} as the input embeddings and are represented as $p = (p_1, \dots, p_{L-1})$ where $p_j \in \mathbb{R}^{d_{model}}$.

2.3.3. Encoder

The encoder layer used in this work is a reduced version of the original transformer encoder [8]. Our architecture (see Figure 2) uses only the multi-head self-attention mechanism along with a residual connection and a layer normalization, which allows to jointly attend to information from different representations subspaces at different positions, performing the attention function in parallel. The reason we do not implement the second sublayer of the original transformer encoder, which is a fully connected feed-forward network, is to reduce the number of model parameters and to avoid the overfitting produced by the small size of the database used in our experiments (see section 3.2). Therefore, when using the multi-head self-attention mechanism each attention head operates on

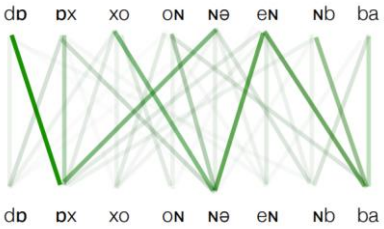


Figure 3: *Behavior of the attention mechanism given a phonetic sequence.*

the input sequence y and computes a new sequence $z = (z_{1:2}, z_{2:3}, \dots, z_{L-1:L})$ where $z_i \in \mathbb{R}^{d_{model}}$.

This final sequence is obtained using the attention-mechanism allowing each token in the input sequence to attend to the other tokens to find those phonetic units that are most important to learn each representation, allowing modeling short and long-distance dependencies between them. We attempt to take advantage of this ability to attend to the most discriminative phonotactic units that could be representative for each language. An example of the attention mechanism applied to a phonetic sequence is shown in Figure 3.

2.3.4. Output

Finally, to get a single representation from the final sequence of embeddings z we choose the average-pooling operation. We apply this operation due to its benefits for tasks with long input sequences [15]. Here we experimented with other techniques such as max-pooling, max attention and the use of the classification token (CLS). However, we obtained better results using average pooling. Next, we use a classification layer with a SoftMax activation given the generated sentence embedding.

3. Database and Experimental Setup

In this section we describe the database used in our experiments, the model parameters and the metric used to measure our performance.

3.1. Database

For our experiments, we used the Kalaka-3 database [16]. This database contains clean and noisy audio recordings of 6 different languages in the closed-set condition (i.e. Basque, Catalan, English, Galician, Portuguese, Spanish), including 108 hours of speech in total. Some relevant statistics of the train, test and validation sets are shown in Table 1.

Table 1: *Statistics of the Kalaka-3 Database*

		Train	Dev	Eval
Languages	Basque	794	70	150
	Catalan	649	79	158
	English	587	81	156
	Galician	975	67	160
	Portuguese	853	84	163
	Spanish	798	77	154
Structure	Nº Files	4656	458	941
	Nº of clean files	3060	-	-
	Nº of noisy files	1596	-	-
	Length <= 30s	2855	121	267
	Length 120s	1801	337	674

As explained before we apply a phoneme recognizer to the audio files to get phoneme sequences, which are later used to form phonetic units. To train our model, we set the number of phonetic units in each sequence to 512 tokens. If any sequence has more, we split it in multiple parts in order to make more training data. We perform this operation only for training, during evaluation we just use the first 512 tokens.

3.2. Experimental Setup

To train our model we attempted to reduce the number of parameters used by the architecture as much as possible due to the small size of the database. In all our experiments, we used a single encoder layer that uses an embedding of size 32 throughout all the operations in the network and a multi-head attention with 2 heads. The phonetic embeddings are then passed to the average-pooling operation to get the final sentence embedding. Finally, this sentence embedding is passed to a final classification layer to obtain per-class probability distributions. We train our model using the Adam Optimizer [17] with a custom learning rate scheduler according to the formula proposed in the original transformer architecture, using as parameters $\beta_1 = 0.9, B_2 = 0.98, \varepsilon = 10^{-9}$. The batch size is set to 64 for all experiments. We train for 25 epochs and select the best model according to the best C_{avg} value. We select this metric to evaluate our model and it is described in the 3.3 subsection [18].

3.3. Average Detection Cost Function

The performance of our experiments is evaluated using accuracy and the average detection cost function (C_{avg}). The latter, is an evaluation metric proposed by NIST that weights the number of false acceptances and false alarms generated by the network, representing them as a detection cost function

$$C_{avg} = \frac{1}{N_L} \cdot \sum_{L_T} \begin{cases} C_{miss} \cdot P_{Target} \cdot P_{miss}(L_T) \\ + \sum_{L_N} C_{FA} \cdot P_{Non-Target} \cdot P_{FA}(L_T, L_N) \\ + C_{FA} \cdot P_{Out-of-Set} \cdot P_{FA}(L_T, L_O) \end{cases} \quad (1)$$

Where:

N_L is the number of languages in the (closed-set) test, L_T is the target language, L_N is the non-target language, L_O is the Out-of-Set "language", $C_{miss} = C_{FA} = 1$, $P_{Target} = 0.5$,

$P_{Out-of-Set} = 0.0$ $(for\ the\ plenty\ closed\ condition)$,

$$P_{NONtarget} = \frac{1 - P_{Target} - P_{out-of-set}}{(N_L - 1)} \quad (2)$$

4. Results

4.1. Experimental Results

The results are presented in Table 2. Here, we show the performance of our model for the development and evaluation sets using different sizes for the phonetic units (i.e. n-gram order). Unfortunately, increasing the n-gram order increases the number of combinations and therefore the number of phonetic units in the corpus; using 2-grams and 3-grams we get a vocabulary of around 10,000 and 200,000 respectively. For this reason, we fixed the vocabulary size used for the word tokenizer. The table shows the best performance we got by using 3-gram phonetic units and setting the vocabulary size to the 30k most common units. We think that this best

performance is due to a better modeling of the local context and that the model is capable of finding more discriminative combinations. On the other hand, we found that when using higher order n-grams there were scattering issues, the training time was increased, and we did not get better performance.

Table 2: *Results*

Size	#Tokens	Dev		Eval	
		Accuracy	Cavg	Accuracy	Cavg
2-gram	5000	85.8	8.6	82.1	10.6
3-gram	30000	**86.0**	**8.4**	**82.7**	**10.2**

In our initial tests, we found that using the original set of BUT Hungarian phonemes performed worse than using the Allosaurus phoneme set. Therefore, all our experiments use the latter. In future work, we will carry additional experiments to better evaluate the reason for this and the possibility of combining information from both sets.

4.2. Baselines

In this subsection we describe the baseline systems which are used to compare the performance of our approach.

4.2.1. RNNs-based architecture using neural phone embeddings

This approach [2] uses phonetic sequences obtained by the Brno recognizer using a PPLRM system, which is followed by a Recurrent neural network that is used to learn phonetic-based embeddings. These phonetic-representations are then used to compare with trained RNN Language Models to obtain entropy values for each language which are used as input to a multi-class logistic classifier. This architecture also uses vocabulary reduction techniques in order to replace rare phonetic units by others that are more discriminative for each language. The best result given in this work uses a size of 3-grams for the phonetic units and k-means clustering as the technique used to reduce the vocabulary.

4.2.2. I-Vectors-Based architectures using Skip-Gram and Glove Models

This work [5] proposed an i-Vector based architecture that uses as input phonetic-based embeddings learned using Skip-Gram and Glove Models. This architecture uses a Multiple Vector Embedding approach (MVE), which consists of using phone-based embeddings trained for each language individually (Language Phone-Based Embeddings) to obtain i-Vectors for each language. Then, these i-Vectors are used as input to a multi-class logistic regression classifier to perform LRE. Finally, the scores provided by each individual language-dependent system are fused to obtain a better performance.

The phoneme sequences used in this approach were obtained by the Brno Recognizer using the Hungarian vocabulary, and its best results are given using a size of 2-grams for the phonetic units

4.3. Comparison to Baseline Algorithms

The comparison of our performance with the best results of our baselines is shown in Table 3. Our approach outperforms all the baselines given an improvement of 45% and 38.5% to i-Vectors based architectures that use the Skip-Gram and Glove Models respectively, and an improvement of 5.5% to the RNN-based architecture.

Table 3: *Comparison of Performance for eval set*

Systems	Cavg	Improvement
Transformer-Based Encoder	**10.2**	(Ours)
i-Vector approach using Skip-gram	18.7	45%
i-Vector approach using Glove	16.7	38.5%
RNN-based architecture	10.8	5.5%

5. Conclusions and Future Work

In this work we show how using a reduced version of the transformer architecture can be beneficial for the LRE task when using phonotactic information. We took advantage of the attention-mechanism used by the transformer-based encoder to find discriminative combinations of phonetic units that helps the model to recognize a language. The model tackles the main hurdle of this task which is the overlapping and high-cooccurrences of phonetic units between languages by learning phonotactic embeddings through attention, attending to the most important tokens in the sequence to find these useful short and long-distance combinations. Our approach showed that it can outperform RNN's and i-Vector based architectures that employ Skip-gram and Glove models. For future work, we plan to apply this model to larger databases. We will also evaluate our approach using longer sequences as input to the model by using different attention windows (as the one used by recent approaches such as the longformer [19] or bigbird [9]) and to use sub-word tokenizers to take advantage of all the vocabulary in the corpus.

6. Acknowledgements

This work has been supported by the Spanish projects AMIC (MINECO, TIN2017-85854-C4-4-R) and CAVIAR (MINECO, TEC2017-84593-C2-1-R) projects partially funded by the European Union. We also gratefully acknowledge the support of the Universidad Politécnica Salesiana.

7. References

[1] M. Zissman, "Comparision of four approaches to automatic language identification of telephone speech", *IEEE Translations on Speech and Audio Processing*, pp. 31-35, 1996.

[2] C. Salamea, L.F. D'Haro, and R. de Córdoba, "Language Recognition Using Neural Phone Embeddings and RNNLMs", *IEEE Latin America Transactions*, vol. 16, no. 7, pp. 2033–2039, 2018.

[3] C. Salamea, L.F D'Haro, R. de Córdoba, and R.S. Segundo, "On the use of phone-gram units in recurrent neural networks for language identification", *in Odyssey*, pp. 117-118.

[4] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. A Field Guide to Dynamical Recurrent Neural Networks", *IEEE*, pp. 237–243, 2001.

[5] C. Salamea, R. de Córdoba, L.F. D'Haro, R.S. Segundo, and J. Ferreiros, "On the use of Phone-based Embeddings for Language Recognition" in *IBERSPEECH 2018 – November 21-23, Barcelona, Spain, Proceedings*, 2018, pp. 55–59.

[6] D. Guthrie, B. Allison, W. Liu, L. Guthrie and Y. Wilks, "A closer look at Skip-Gram modelling", *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 1–4, 2006.

[7] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representations", *Proceedings of conference on empirical methods in natural language processing*, pp. 1532–1543, 2014.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need",

in Advances in Neural Information processing systems, pp. 5998-6008, 2017.

[9] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang and A. Ahmed, "Big Bird: Transformers for Longer Sequences", *in NIPS – 2020*.

[10] C. Sun, L. Huan, and X. Qiu, "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computer Linguistics: Human Language Technologies*, vol.1., pp. 380-385, 2019.

[11] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of bidirectional transformers for language understanding", *in NAACL-HLT*, arXiv: 1810.04805.

[12] S. Hochreiter, and J. Schmidhuber, "Long short-term memory", *in Neural computation*, 9(8), pp. 1735-1780, 1997.

[13] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos et al, "Universal Phone Recognition with a Multilingual Allophone System", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8249–8253, 2020.

[14] P. Schwarz, "Phoneme Recognition based on Long Temporal Context, PhD Thesis", *Brno University of Technology*, 2009

[15] P. Maini, K. Kolluru, D. Pruti, Mausam, "Why and when should you pool? Analyzing Pooling in Recurrent Architectures", *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4568-4586.

[16] L. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, and G. Bordel, "KALAKA-3: a database for the assessment of spoken language recognition technology on Youtube audios", *in Language Resources and Evaluation*, pp. 221-243, 2016.

[17] P. Diederik, Kingma and Jimmy Ba, "Adam: A method for stochastic optimization", 2017. arXiv: 1412.6980

[18] A. Martin and C. Greenberg, "The 2009 NIST Language Recognition Evaluation", *in Speaker and Language Recognition Workshop, IEEE Odyssey*, pp. 165-171,2010.

[19] I. Beltagy, M.E. Petters and A. Cohan, 2020. Longformer: The long-document transformer. a*rXiv preprint* arXiv:2004.05150

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Adversarial Transformation of Spoofing Attacks for Voice Biometrics

Alejandro Gomez-Alanis, Jose A. Gonzalez-Lopez, Antonio M. Peinado

University of Granada

agomezalanis@ugr.es, joseangl@ugr.es, amp@ugr.es

Abstract

Voice biometric systems based on automatic speaker verification (ASV) are exposed to *spoofing* attacks which may compromise their security. To increase the robustness against such attacks, anti-spoofing or presentation attack detection (PAD) systems have been proposed for the detection of replay, synthesis and voice conversion based attacks. Recently, the scientific community has shown that PAD systems are also vulnerable to adversarial attacks. However, to the best of our knowledge, no previous work have studied the robustness of full voice biometrics systems (ASV + PAD) to these new types of adversarial *spoofing* attacks. In this work, we develop a new adversarial biometrics transformation network (ABTN) which jointly processes the loss of the PAD and ASV systems in order to generate white-box and black-box adversarial *spoofing* attacks. The core idea of this system is to generate adversarial *spoofing* attacks which are able to fool the PAD system without being detected by the ASV system. The experiments were carried out on the ASVspoof 2019 corpus, including both logical access (LA) and physical access (PA) scenarios. The experimental results show that the proposed ABTN clearly outperforms some well-known adversarial techniques in both white-box and black-box attack scenarios.

Index Terms: Adversarial attacks, automatic speaker verification (ASV), presentation attack detection (PAD), voice biometrics.

1. Introduction

Voice biometrics aims to authenticate the identity claimed by a given individual based on the speech samples measured from his/her voice. Automatic speaker verification (ASV) [1] is the conventional way to put voice biometrics into practical usage. However, in recent years, ASV technology has been shown to be at risk of security threats performed by impostors who want to gain fraudulent access by presenting speech resembling the voice of a legitimate user [2, 3]. Impostors could use either logical access (LA) attacks [4], such as text-to-speech synthesis (TTS) and voice conversion (VC) based attacks, or physical access (PA) attacks such as replay based attacks [5].

To protect voice biometrics systems [6], it is common to develop anti-spoofing or presentation attack detection (PAD) [7] techniques which allow for differentiating between *bonafide* and *spoofing* speech [8, 9, 10]. Typically, the resulting biometrics system is a score-level cascaded integration of PAD and ASV subsystems, as depicted in Fig. 1. This is the same integration as the one used in the last two ASVspoof challenges [5, 11].

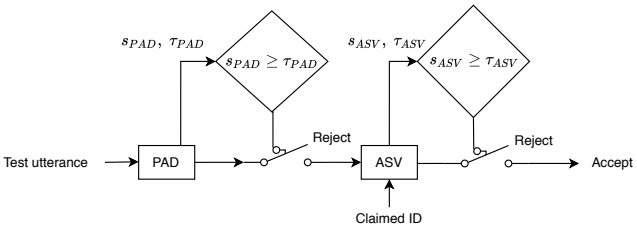


Figure 1: *Block diagram of a score-level cascaded integration biometrics system. s_{PAD}, τ_{PAD} and s_{ASV}, τ_{ASV} denote the scores and thresholds of the PAD and ASV systems, respectively.*

To make things more complex, different investigations [12, 13] have recently shown that PAD systems are also vulnerable to adversarial attacks [14]. These attacks can easily fool deep neural network (DNN) models by perturbing benign samples in a way normally imperceptible to humans [15]. Adversarial attacks can be divided into two main categories: white-box and black-box attacks. In this work, we refer to white-box attacks as those where the attacker can access all the information of the victim model (i.e., model architecture and its weights). Likewise, we will use the term black-box for those attacks where the attacker does not know any information about the victim model but it can be queried multiple times in order to estimate a surrogate model (student) of the victim model (teacher), using the binary responses (acceptance/rejection) of the victim model as ground-truth labels.

The main contributions of this work are:

- Investigate the robustness of full voice biometrics systems (ASV + PAD) under the presence of adversarial *spoofing* attacks.
- Propose an adversarial biometrics transformation network (ABTN) which is able to generate adversarial *spoofing* attacks in order to fool the PAD system without being detected by the ASV system.
- To the best of our knowledge, adversarial *spoofing* attacks have only been studied on logical access scenarios (TTS and VC based attacks). In this work, we also include physical access scenarios (replay based attacks).

The rest of this paper is organized as follows. Section 2 outlines some well-known adversarial attacks employed as baselines in this work. Then, in Section 3, we describe the proposed ABTN for white-box and black-box scenarios. After that, Section 4 outlines the speech corpora, systems details, and metrics employed in the experiments. Section 5 discusses the experimental results. Finally, we summarize the conclusions derived from this research in Section 6.

This work has been supported by the Spanish Ministry of Science and Innovation Project No. PID2019-104206GB-I00/AEI/10.13039/501100011033. Alejandro Gomez-Alanis holds a FPU fellowship from the Spanish Ministry of Education (FPU16/05490). Jose A. Gonzalez-Lopez holds a Juan de la Cierva-Incorporación fellowship from the Spanish Ministry of Science, Innovation and Universities (IJCI-2017-32926). We also acknowledge the support of NVIDIA Corporation with the donation of a Titan X GPU.

10.21437/IberSPEECH.2021-54

2. Background

Adversarial *spoofing* examples can be generated by adding a minimally perceptible perturbation to the input *spoofing* utterance in order to do a refinement of the *spoofing* attack. In this work, we focus on targeted attacks, which aim to fool the PAD system by maximizing the probability of a targeted class (*bonafide*) different from the correct class (*spoof*). Specifically, to generate adversarial *spoofing* attacks, we fix the parameters $\boldsymbol{\theta}$ of a well-trained DNN-based PAD model and perform gradient descent to update the *spoofing* spectra of the input utterance so that the PAD model classifies it as a *bonafide* utterance. Mathematically, our goal is to find a sufficiently small perturbation $\boldsymbol{\delta}$ which satisfies:

$$\begin{aligned} \tilde{\boldsymbol{X}} &= \boldsymbol{X} + \boldsymbol{\delta}, \\ f_{\boldsymbol{\theta}}(\boldsymbol{X}) &= y, \\ f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}) &= \tilde{y}, \end{aligned} \tag{1}$$

where f is a well-trained DNN-based PAD model parameterized by $\boldsymbol{\theta}$, $\boldsymbol{X}$ denotes the sequence of speech feature vectors extracted from the input *spoofing* utterance (short time Fourier transform (STFT), typically), y is the true label corresponding to $\boldsymbol{X}$, $\tilde{y}$ is the targeted label class of the attack (*bonafide* class), $\tilde{\boldsymbol{X}}$ denotes the perturbed input features, and $\boldsymbol{\delta}$ is the additive perturbation. Typically, $\boldsymbol{\Delta}$ is the feasible set of the allowed perturbation $\boldsymbol{\delta}$ ($\boldsymbol{\delta} \in \boldsymbol{\Delta}$), which formalizes the manipulative power of the adversarial attack. Normally, $\boldsymbol{\Delta}$ is a small l_∞-norm ball, that is, $\boldsymbol{\Delta} = \{\boldsymbol{\delta} \mid \|\boldsymbol{\delta}\|_\infty \leq \epsilon\}, \epsilon \geq 0 \in \mathbb{R}$.

There are multiple ways to generate the perturbation $\boldsymbol{\delta}$, where the fast gradient sign method (FGSM) [16] and the projected gradient descent (PGD) [17] methods are the most popular adversarial attack procedures. The FGSM attack consists of taking a single step along the direction of the gradient, i.e.,

$$\boldsymbol{\delta} = \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{X}} Loss(\boldsymbol{\theta}, \boldsymbol{X}, y)), \tag{2}$$

where $Loss$ denotes the loss function of the neural network ($\boldsymbol{\theta}$), and the sign method simply takes the sign of its gradient. Unlike the FGSM, which is a single-step method, the PGD is an iterative method. Starting from the original input utterance $\boldsymbol{X}_0 = \boldsymbol{X}$, the input utterance is iteratively updated as follows:

$$\begin{aligned} &\boldsymbol{X}_{n+1} = \text{clip}(\boldsymbol{X}_n + \alpha \cdot \text{sign}(\nabla_{\boldsymbol{X}} Loss(\boldsymbol{\theta}, \boldsymbol{X}, y)), \\ &\text{for } n = 0, ..., N-1, \end{aligned} \tag{3}$$

where $n = 0, ..., N-1$ is the iteration index, N is the number of iterations, $\alpha = \epsilon/N$, and the clip() function applies element-wise clipping such that $\|\boldsymbol{X}_n - \boldsymbol{X}\|_\infty \leq \epsilon, \epsilon \geq 0 \in \mathbb{R}$.

3. Proposed method

The performance of the FGSM and PGD methods is limited by the possibility of sticking at local optima of the loss function. Moreover, both methods have a limited search space ($\boldsymbol{\Delta}$) so that the perturbed *spoofing* speech $\tilde{\boldsymbol{X}}$ is perceptually indistinguishable from the original *spoofing* speech $\boldsymbol{X}$.

In this work, we propose the Adversarial Biometrics Transformation Network (ABTN), which is a neural network that transforms a *spoofing* speech signal into an adversarial *spoofing* speech signal against a target biometrics system. Formally, an ABTN can be defined as a neural network $g_{f,h} : \boldsymbol{X} \rightarrow \tilde{\boldsymbol{X}}$, where $f(\boldsymbol{X})$ and $h(\boldsymbol{X})$ are the PAD and ASV models of the target biometrics system, respectively. The PAD and ASV models can provide either a probability distribution across class labels (white-box scenario) or just a binary decision (black-box scenario). In both scenarios, the objective of the ABTN is to generate adversarial *spoofing* attacks from *spoofing* speech in order to fool the PAD system while not being detected by the ASV system, i.e., while not modifying the speaker information.

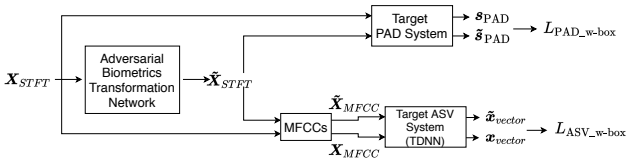


Figure 2: *Proposed adversarial biometrics transformation system for white-box scenarios.*

3.1. White-box scenario

The architecture of the proposed ABTN system for the white-box scenario is depicted in Fig. 2. The output of the ABTN is fed into the target biometrics system which is composed of a PAD and an ASV system based on a time-delay neural network (TDNN) [18] for x-vector extraction (in fact, this is the only component of the ASV system that we need). The objective of this system is to train the ABTN so that it can generate adversarial attacks from *spoofing* speech which are able to fool the PAD system while, at the same time, it does not cause any changes to the ASV output (i.e., it does not change the speaker representation given by the corresponding x-vector). To train the ABTN, the PAD and ASV network parameters are frozen but the gradients are computed along them in order to back-propagate them to the ABTN parameters. To find the optimal parameters of the ABTN in the white-box (w-box) scenario, we minimize the following loss function:

$$L_{\text{w-box}} = L_{\text{PAD\_w-box}}(\boldsymbol{s}_{\text{PAD}}, \tilde{\boldsymbol{s}}_{\text{PAD}}) + \beta \cdot L_{\text{ASV\_w-box}}(\boldsymbol{x}_{\text{vector}}, \tilde{\boldsymbol{x}}_{\text{vector}}), \tag{4}$$

where,

$$L_{\text{PAD\_w-box}}(\boldsymbol{s}_{\text{PAD}}, \tilde{\boldsymbol{s}}_{\text{PAD}}) = \|r_\alpha(\boldsymbol{s}_{\text{PAD}}) - \tilde{\boldsymbol{s}}_{\text{PAD}}\|_2, \tag{5}$$

$$L_{\text{ASV\_w-box}}(\boldsymbol{x}_{\text{vector}}, \tilde{\boldsymbol{x}}_{\text{vector}}) = \|\boldsymbol{x}_{\text{vector}} - \tilde{\boldsymbol{x}}_{\text{vector}}\|_2. \tag{6}$$

$L_{\text{PAD\_w-box}}$ and $L_{\text{ASV\_w-box}}$ are the loss components associated to the PAD and ASV systems, respectively, and β is a hyper-parameter to weight the importance of the two losses. $\boldsymbol{s}_{\text{PAD}}$ and $\tilde{\boldsymbol{s}}_{\text{PAD}}$ are the probability output vectors from the PAD system of the original and adversarial *spoofing* utterances, respectively. Likewise, $\boldsymbol{x}_{\text{vector}}$ and $\tilde{\boldsymbol{x}}_{\text{vector}}$ denote the x-vectors of the original and adversarial *spoofing* utterances, respectively, and r_α is a reranking function which can be formulated as

$$r_\alpha(\boldsymbol{s}_{\text{PAD}}) = norm\left(\begin{cases} \alpha \cdot \max(\boldsymbol{s}_{\text{PAD}}) & k = 0 \\ \boldsymbol{s}_{\text{PAD}}(k) & k \neq 0 \end{cases}\right), \tag{7}$$

where k is the index class variable of the $\boldsymbol{s}_{\text{PAD}}$ probability vector, $\alpha > 1$ is an additional hyper-parameter which defines how large $\boldsymbol{s}_{\text{PAD}}(k = 0)$, i.e., the probability of the *bonafide* class, is with respect to the current maximum probability class, and $norm$ is a normalizing function which rescales its input to be a valid probability distribution.

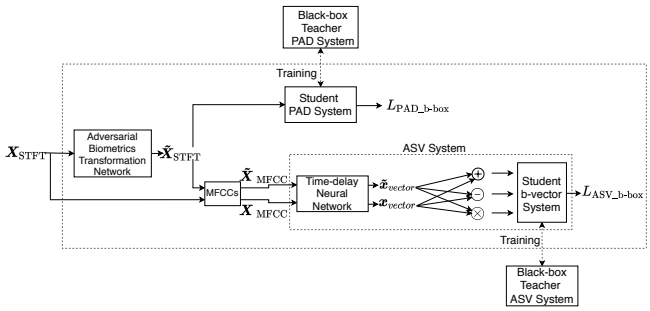


Figure 3: *Proposed adversarial biometrics transformation system for black-box scenarios.*

3.2. Black-box scenario

The architecture of the proposed ABTN system for the black-box scenario is depicted in Fig. 3. Similarly to the white-box scenario, the objective of this system is to generate adversarial attacks from *spoofing* speech which are able to fool the target (teacher) PAD system and, at the same time, bypass the target (teacher) ASV system by not modifying the speaker information represented by the corresponding x-vector. However, the limitation of the black-box scenario is that we do not have access to the parameters of the target biometrics system. Thus, we train a student PAD and a b-vector [19] based ASV systems by making requests to the target black-box biometrics system which only responds with a binary decision of acceptance or rejection, using these binary decisions as ground-truth labels. Therefore, the student PAD and b-vector systems are trained as binary classifiers in order to mimick the performance of the teacher PAD and ASV systems, respectively. Specifically, the student b-vector system computes the probability that the two input x-vectors belong to the same speaker, i.e., that $P(b(\boldsymbol{x}_{\text{vector}}, \tilde{\boldsymbol{x}}_{\text{vector}}) = 1)$, where b denotes the b-vector model.

To train the ABTN in the black-box scenario, the student PAD and ASV network parameters are also frozen but the gradients are computed along them in order to back-propagate them to the ABTN parameters. To find the optimal parameters of the ABTN in the black-box (b-box) scenario, we minimize the following loss function:

$$L_{\text{b-box}} = L_{\text{PAD\_b-box}}(\tilde{\boldsymbol{s}}_{\text{PAD}}) + \beta \cdot L_{\text{ASV\_b-box}}(\boldsymbol{x}_{\text{vector}}, \tilde{\boldsymbol{x}}_{\text{vector}}), \quad (8)$$

where,

$$L_{\text{PAD\_b-box}}(\tilde{\boldsymbol{s}}_{\text{PAD}}) = \|onehot(k=0) - \tilde{\boldsymbol{s}}_{\text{PAD}}\|_2 , \quad (9)$$

$$L_{\text{ASV\_b-box}}(\boldsymbol{x}_{\text{vector}}, \tilde{\boldsymbol{x}}_{\text{vector}}) = 1 - P(b(\boldsymbol{x}_{\text{vector}}, \tilde{\boldsymbol{x}}_{\text{vector}}) = 1). \quad (10)$$

$L_{\text{PAD\_b-box}}$ and $L_{\text{ASV\_b-box}}$ are the loss components associated to the PAD and ASV systems, respectively. Moreover, the function $onehot$ denotes the one-hot function and $k = 0$ is the index of the *bonafide* class, so that the PAD system is fooled by firing the input *spoofing* utterance as a *bonafide* utterance.

4. Experimental Setup

This section briefly describes the speech corpora and metrics employed in our experiments, as well as the details of the proposed system.

4.1. Speech corpora

We conducted experiments on the ASVspoof 2019 database [20] which is split into two partitions for the assessment of LA and PA scenarios. This database also includes protocols for evaluating the performance of PAD, ASV and integration (biometrics) systems. Thus, we used this corpus for training the standalone PAD systems in the LA and PA scenarios, separately. Then, we generated adversarial *spoofing* attacks using only the *spoofing* utterances, so that they can bypass the biometrics system. We did not generate any adversarial examples from *bonafide* utterances, since we argue that they would not be *bonafide* anymore.

On the other hand, we also employed the Voxceleb1 [21] to train a TDNN [18] as an x-vector extractor for the ASV system. Also, following [6], a b-vector [19] ASV scoring system was trained in the black-box scenario using the *bonafide* utterances from the ASVspoof 2019 and Voxceleb1 development datasets.

4.2. Spectral analysis

Speech signals were analyzed using a Hanning analysis windows of 25 ms length with 10 ms of frame shift. Log-power magnitude spectrum features (STFT) with 256 frequency bins were obtained to feed all the PAD systems. The ASV systems were fed with Mel-frequency cepstral coefficients (MFCCs) obtained with the Kaldi recipe [22]. Only the first 600 frames of each utterance were used to extract acoustic features.

4.3. Implementation details

Two state-of-the-art PAD systems were adapted from different works, i.e., a light convolutional neural network (LCNN) [2] and a Squeeze-Excitation network (SENet50) [23]. The PAD scores were directly obtained from the *bonafide* class of the softmax output. For ASV, a TDNN x-vector model [18] was trained as an embedding extractor. Then, a probabilistic linear discriminant analysis (PLDA) [24] and a b-vector system [19] were trained as ASV scoring systems.

The proposed ABTN is formed by five convolutional layers with 16, 32, 48, 48 and 3 channels, respectively, and a kernel size of 3×3, followed by leaky ReLU activations. It was trained using the Adam optimizer [25] with a learning rate of $3 \cdot 10^{-4}$. Also, early stopping was applied to stop the training process when no improvement of the loss across the validation set was obtained. The values of α and β were empirically set to 10 and 0.001, respectively, using a grid search on the validation set.

4.4. Evaluation setup

The PAD systems were evaluated using the pooled equal error rate ($\text{EER}_{\text{spoof}}$) across all attacks. Likewise, the ASV systems were also evaluated using the EER_{ASV}, employing both *bonafide* utterances (target and non-target) and *spoofing* utterances. Any utterance rejected by either the PAD or ASV subsystems was assigned arbitrarily a $-\infty$ score for computing the integration performance. Then, the integration (biometrics) systems were evaluated using the joint EER ($\text{EER}_{\text{joint}}$) and the minimum normalized detection cost function (min-tDCF) [26] with the same configuration as the one employed in the ASVspoof 2019 challenge [11]. All the PAD, ASV and biometrics systems were evaluated using the ASVspoof 2019 test datasets.

System	Logical Access Attacks				Physical Access Attacks			
	$EER_{spoof}(\%)$	$EER_{ASV}(\%)$	$EER_{joint}(\%)$	min-tDCF	$EER_{spoof}(\%)$	$EER_{ASV}(\%)$	$EER_{joint}(\%)$	min-tDCF
No Attack	5.91	31.10*	20.13	0.1252	4.77	18.62*	13.37	0.1238
FGSM ($\epsilon = 0.1$)	5.98	31.14*	20.32	0.1279	7.50	18.65*	15.47	0.2157
PGD ($\epsilon = 0.1$)	5.95	31.13*	20.25	0.1267	6.08	18.63*	14.38	0.1717
FGSM ($\epsilon = 1.0$)	8.15	31.53*	25.44	0.1287	35.64	18.71*	26.54	0.9335
PGD ($\epsilon = 1.0$)	7.02	31.46*	25.37	0.1266	44.42	18.83*	26.77	0.9665
FGSM ($\epsilon = 2.0$)	2.01	30.11*	14.13	0.0623	1.02	17.61*	11.82	0.0380
PGD ($\epsilon = 2.0$)	4.97	31.38*	22.62	0.1078	29.29	18.44*	25.28	0.8677
FGSM ($\epsilon = 5.0$)	0.00	19.46*	2.45	0.0000	0.00	11.37*	11.79	0.0000
PGD ($\epsilon = 5.0$)	0.16	19.09*	2.56	0.0058	0.00	9.48*	11.79	0.0000
Proposed ABTN	**35.19**	**31.52***	**39.15**	**0.5829**	**95.17**	**18.87***	**36.63**	**1.0000**

Table 1: *Results of the black-box adversarial attacks on the ASVspoof 2019 logical access (LA) and physical access (PA) test sets in terms of $EER_{spoof}(\%)$, $EER_{ASV}(\%)$, $EER_{joint}(\%)$ and min-tDCF. The target PAD system is based on a LCNN, while the student PAD system is based on a SENet50. The target ASV system is based on a TDNN + PLDA, while the student ASV system is based on a TDNN + b-vector. (*) The ASV evaluation includes both bonafide and spoofing utterances.*

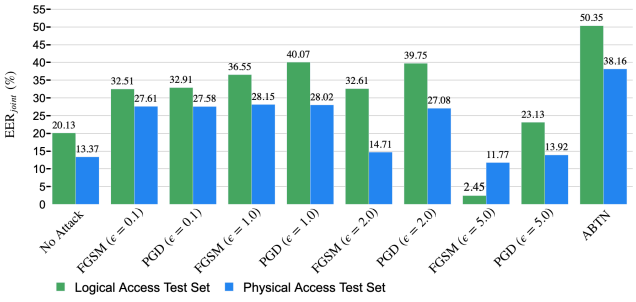


Figure 4: *$EER_{joint}(\%)$ of the white-box adversarial attacks on the ASVspoof 2019 logical and physical access test sets.*

5. Experimental Results

The performance of the baseline biometrics system is shown in Table 1 as 'No Attack'. The LA and PA PAD systems are among the best single systems evaluated in the ASVspoof 2019 challenge [11]. The ASV system yields an EER of 4.75 and 7.25% in the LA and PA datasets when evaluating only the target and non-target *bonafide* utterances. However, its performance is degraded to 31.10 and 18.62% in the LA and PA test datasets when the *spoofing* utterances are also evaluated, as shown in Table 1.

5.1. White-box scenario

Fig. 4 shows the EER_{joint} of the white-box adversarial attacks evaluated in the ASVspoof 2019 LA and PA test sets. The PAD and ASV systems are the state-of-the-art LCNN and TDNN + PLDA, respectively. As it was expected, PGD achieves slightly better results than FGSM due to its iterative procedure for generating the adversarial attacks. Moreover, the proposed ABTN outperforms the rest of adversarial attacks, obtaining 10.28% and 10.14% higher EER_{joint} with respect to the best PGD configuration ($\epsilon = 1.0$) in the LA and PA test sets, respectively. It is worth noticing that when the hyper-parameter ϵ of the FGSM and PGD methods is equal or higher than 2.0, the biometrics system is able to detect the perturbation noise added by these adversarial attacks. In these cases, the performance of the *spoofing* attacks is even worse than when not using any adversarial attack (denoted by 'No Attack').

5.2. Black-box scenario

Table 1 shows the performance metrics for the black-box scenario. The target biometrics system consists of the same state-of-the-art LCNN (PAD) and TDNN + PLDA (ASV) systems evaluated in the previous section. The student PAD and ASV systems are the SENet50 and the TDNN + b-vector systems, respectively.

The proposed ABTN attacks outperform the best FGSM and PGD configurations by 27.04 and 50.75% of EER_{spoof}, and by 13.71 and 9.86% of EER_{joint}, respectively. Also, the min-tDCF metric, which shows the performance of the biometrics system on a different operating point with respect to the EER_{joint} [26], is significatively higher for the proposed ABTN adversarial attacks. As in the white-box scenario, it is worth noticing that the best adversarial attacks do not affect the performance of the ASV system with respect to the baseline system, since the perturbation noise of these attacks is not detected by the ASV system. However, when the hyper-parameter $\epsilon \geq 2.0$, both the PAD and ASV systems are able to detect the perturbations added by the FGSM and PGD methods, and hence, the biometrics system performs even better than the baseline system (denoted by 'No Attack'). However, the proposed ABTN method does not suffer from this issue since it is trained so that the added perturbation noise does not modify the speaker information from the *spoofing* utterance.

6. Conclusion

In this work, we studied the robustness of state-of-the-art voice biometrics systems (ASV + PAD) under the presence of adversarial *spoofing* attacks. Moreover, we proposed an adversarial biometrics transformation network (ABTN) for both white-box and black-box scenarios which is able to generate adversarial *spoofing* attacks in order to fool the PAD system without being detected by the ASV system. Experimental results have shown that biometric systems are highly sensitive to adversarial *spoofing* attacks in both logical and physical access scenarios. Moreover, the proposed ABTN system clearly outperforms other popular adversarial attacks such as the FGSM and PGD methods in both white-box and black-box scenarios. In the future, we would like to use the generated adversarial attacks for adversarial training in order to make the biometrics system more robust against these attacks.

7. References

[1] R. Naika, "An overview of automatic speaker verification system," in *Advances in Intelligent Systems and Computing*. New York, NY, USA: Springer, 2018, vol. 673.

[2] A. Gomez-Alanis, J. A. Gonzalez-Lopez, and A. M. Peinado, "A kernel density estimation based loss function and its application to ASV-spoofing detection," *IEEE Access*, vol. 8, pp. 108 530–108 543, 2020.

[3] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A gated recurrent convolutional neural network for robust spoofing detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 1985–1999, 2019.

[4] Z. Wu, T. Kinnunen, N. W. D. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 2037–2041.

[5] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. W. D. Evans, J. Yamagishi, and K.-A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 2–6.

[6] A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. P. Dubagunta, A. M. Peinado, and M. Magimai.-Doss, "On joint optimization of automatic speaker verification and anti-spoofing in the embedding space," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1579–1593, 2021.

[7] "Presentation attack detection." [Online]. Available: https://www.iso.org/standard/67381.html

[8] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A deep identity representation for noise robust spoofing detection," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 676–680.

[9] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "Performance evaluation of front- and back-end techniques for asv spoofing detection systems based on deep features," in *Proc. Iberspeech*, Barcelona, Spain, 2018, pp. 45–49.

[10] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection," in *Proc. Interspeech*, Graz, Austria, 2019, pp. 1068–1072.

[11] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Interspeech*, Graz, Austria, 2019, pp. 1008–1012.

[12] S. Liu, H. Wu, H. yi Lee, and H. Meng, "Adversarial attacks on spoofing countermeasures of automatic speaker verification," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 312–319.

[13] Y. Zhang, Z. Jiang, J. Villalba, and N. Dehak, "Black-box attacks on spoofing countermeasures using transferability of adversarial examples," in *Proc. Interspeech*, 2020, pp. 4238–4242.

[14] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346 – 360, 2020.

[15] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. International Conference on Learning Representations (ICLR)*, Banf, Alberta, Canada, 2014.

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[17] Y. Deng and L. J. Karam, "Universal adversarial attack via enhanced projected gradient descent," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 1241–1245.

[18] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018, pp. 5329–5333.

[19] H.-S. Lee, Y. Tso, Y.-F. Chang, H.-M. Wang, and S.-K. Jeng, "Speaker verification using kernel-based binary classifiers with binary operation derived features," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 1660–1664.

[20] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, and N. E. et al., "ASVspoof 2019: a large-scale public database of synthetized, converted and replayed speech," *Computer Speech and Language*, p. 101114, 2020.

[21] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 2616–2620.

[22] "SRE16 xvector model." [Online]. Available: http://kaldi-asr.org/models/m3

[23] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual Networks," in *Proc. Interspeech*, 2019, pp. 1013–1017.

[24] S. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[26] T. Kinnunen, K. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. Reynolds, "t-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Proc. Odyssey*, Les Sables d'Olonne, France, 2018, pp. 312–319.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Active correction for speaker diarization with human in the loop

Yevhenii Prokopalo[1], *Meysam Shamsi*[1], *Loïc Barrault*[2], *Sylvain Meignier*[1], *Anthony Larcher*[1]

[1] LIUM, Le Mans Université, [2]The University of Sheffield

firstname.lastname@univ-lemans.fr, l.barrault@sheffield.ac.uk

Abstract

State of the art diarization systems now achieve decent performance but those performances are often not good enough to deploy them without any human supervision. In this paper we propose a framework that solicits a human in the loop to correct the clustering by answering simple questions. After defining the nature of the questions, we propose an algorithm to list those questions and two stopping criteria that are necessary to limit the work load on the human in the loop. Experiments performed on the ALLIES dataset show that a limited interaction with a human expert can lead to considerable improvement of up to 36.5% relative diarization error rate (DER) compared to a strong baseline.

Index Terms: Speaker diarization, Active learning, Clustering

1. Introduction

Speaker diarization answers the question "Who speaks when?" along an audio recording[1, 2]. Being important for audio indexing, it is also a pre-processing step for many speech tasks such as speech recognition, spoken language understanding or speaker recognition. For an audio stream that involves multiple speakers, diarization is usually achieved in two steps: i) a segmentation of the audio stream into segments involving a single acoustic event (speech from one speaker, silence, noise...); ii) a clustering that groups segments along the stream when they belong to the same class of event. A last step could be added to name the resulting speakers but this step is out of the scope of this paper.

Modern diarization systems achieve decent performance depending on the type of data they process [3] but those performances are often not good enough to deploy such systems without any human supervision [4, 5]. Human assisted learning offers a way to achieve better performance by engaging an interaction between the automatic system and a human expert in order to correct or guide the automatic diarization process [6, 7]. Amongst the different modes of human assisted learning, our work focuses on active learning where the automatic system, while processing an incoming stream of audio, is allowed to ask simple questions to the human expert [8].

We propose in this study a system architecture depicted in Figure 1. Given an audio file, the human assisted speaker diarization system (HASDS) first produces an hypothesis based on which a questioning module sends a request to the human expert. The expert's answer is taken into account to correct the hypothesis and possibly adapt the diarization system. This process iterates until reaching a stopping criteria out of those three: (i) the system has no more question (ii) the human expert stops answering (iii) a maximum interaction cost is reached. In this work, we define a binary question that allows a user/system interaction and propose two questioning methods with the associated correction module. The scope of this paper doesn't encompass the system adaptation that will be studied in a future work.

This work has been funded by the CHIST-ERA project ALLIES (ARN-17-CHR2-0004-01)

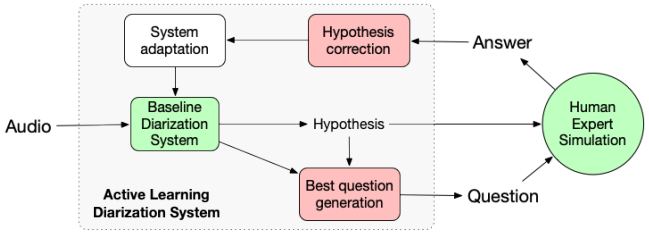


Figure 1: *Life-cycle of a human assisted speaker diarization system.*

Section (2) describes the related works. Section 3 introduces the HASDS, whose evaluation is described and discussed in Section 4. The outcomes and the perspectives of this study are summarized in Section 5.

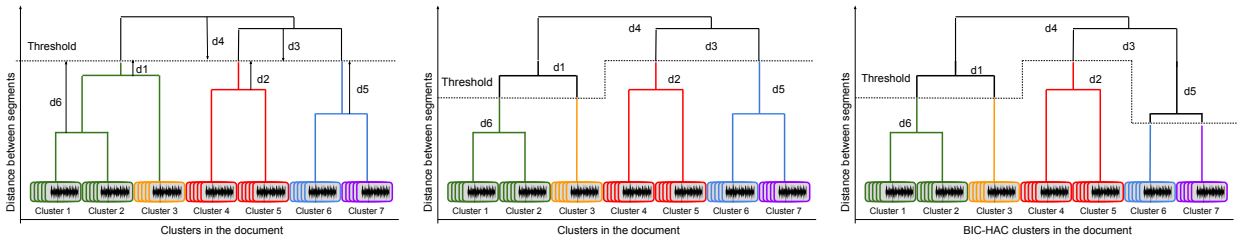
2. Related work

Literature on active learning for speaker diarization is very sparse and existing approaches are complementary to our work more than competitive. In [9], active learning is used to find the initial number of speaker models in a collection of documents. This information is used to perform speaker diarization without involving the human expert anymore. In [10], multi-modal active learning is proposed to process speech segments according to their length to add missing labels, task that is out of the scope of our study. [4] proposed an active learning framework to apply different types of corrections together with metrics to evaluate the cost of human-computer interactions. Unlike previous cited papers, in our work, one interaction with the human expert can lead to correcting a whole cluster of segments (obtained with first of two clustering steps) instead of correcting a single segment only.

In [11], active learning is used to leverage training data and improve a speaker recognition system similar to the one we use for clustering. Active learning based approaches have been developed for other speech processing tasks including speech recognition [12, 13, 14], language recognition[15], speech activity detection [16] or speech emotion recognition [17] but are not directly applicable to speaker diarization.

Active learning literature for clustering is much wider [18, 19] but mostly focuses on K-mean clustering [20, 21] or spectral clustering [22, 23]. Hierarchical agglomerative clustering, that is used in many speaker diarization systems including our baseline, has also been studied for semi-supervised clustering [24, 25].Those studies propose to use predefined constraints to modify the clustering tree. In our work, instead of modifying the dendrogram, we propose a dynamic approach to update the threshold used to merge and split the clusters.

10.21437/IberSPEECH.2021-55



(a) *Initial HAC* (b) *Stopping criteria : 2c criteria* (c) *Stopping criteria : All*

Figure 2: *The change of HAC dendrogram after interaction with human expert*

Regarding evaluation of the active learning process, multiple approaches have been proposed [26, 4]. In [4], systems are evaluated by DER together with an estimate of the human work load to correct the hypotheses. We make the choice to use a penalized version of DER described in our previous work [27]. The human correction effort is computed to be in the same unit and thus added to the DER in order to provide a single performance estimator reflecting both the final performance and the cost of interacting with human.

3. Human Assisted Diarization System

The proposed Human Assisted Speaker Diarization System (HASDS) is depicted in Figure 1 and includes 4 modules. A fully automatic baseline diarization system, a question generation module, a correction module and an adaptation module. This section describes the baseline diarization system that is considered fixed in this work before introducing the proposed question generation and correction modules. The adaptation module is out of the scope of this work and will be considered in future work.

Given an audio stream, diarization consists of producing a segmentation hypothesis, i.e. a list of segments and speaker IDs with each segment allocated to a single speaker ID (segments might overlap). Diarization errors can be due to errors in the segment borders or in a wrong label allocation. The former error being the most harmful in terms of performance [4], this work only focuses on correcting labeling errors.

3.1. Baseline automatic diarization module

To provide a fair study, the chosen baseline system is the best automatic diarization system available at LIUM for the given task. This system performs the diarization in two steps: a segmentation process, that splits the audio stream into (possibly overlapping) segments and a clustering process, that groups the segments into clusters: one cluster per speaker. Since this study only focuses on labeling errors, the segmentation step is considered perfect, i.e., border of the speech segments are taken from the reference. The clustering is performed in four steps: (i) a first hierarchical agglomerative clustering (HAC) is performed on vectors of 13 MFCC using the BIC criteria [28]; (ii) a Viterbi decoding is then used to smooth the segment borders along the audio stream; (iii) x-vectors are extracted from each segment and averaged to provide a single x-vector per BIC-HAC cluster; (iv) a second (final) HAC clustering is done by using x-vectors. The distance matrix used for this clustering is computed using a PLDA scoring [29]. X-vectors are extracted using the SincNet extractor described in [30] and their dimension is 100. A simplified-PLDA [29] is trained using an EigenVoice matrix of rank 100.

Applying two consecutive clustering makes the application of active correction more complex but removing one of the steps degrade the performance of the baseline system, thus we chose to keep the two consecutive clustering but to only apply active correction to the second clustering step while considering the BIC-HAC clusters as frozen. This choice has the advantage to reduce the correction to a simpler HAC-tree correction process. Another drawback is that errors from the BIC-HAC clustering will not be corrected and purity of those clusters is thus very important.

3.2. Question generation module

Assuming that correcting the clustering provides higher gains than segmentation and considering an HAC clustering algorithm, we propose to limit the human/system interaction to a simple binary question that can be asked for each node of the HAC dendrogram (Figure 2a). HAC clustering is done with no prior on the number of clusters and the threshold is empirically determined on a development set. Once this threshold is set, it separates the dendrogram in two parts (above and below the threshold). From this point, the same question can be asked to the human expert for each node of the dendrogram: *"Do the two branches of the node belong to the same speaker?"* . A "yes" answer from the human expert requires either to join the two branches of a node above the threshold (merging operation) or to leave as it is the branches of a node below (no splitting required). In case of a "no" answer, a node above the threshold would not be modified (no merging required) and the two branches of a node below the threshold would be separated (split operation).

One must now determine which node to ask about and when to stop asking. To do so, we rely on the distance between the threshold and the nodes, referred to as *delta* to differentiate with distance between x-vectors. Examples of those *delta* are labeled $d1$ to $d6$ on Figure 2a. Nodes are ranked in increasing order according to their absolute *delta* value. We propose to ask questions about the nodes in this order, and consider two different stopping criteria. First, a **Two confirmation criteria (*2c criteria*)** illustrated in Figure 2b, in which we assume that if a node above the threshold is confirmed by the human expert to be separated ("no" answer) then other nodes above it, with higher *deltas* will not be investigated. Similarly, if one node below the threshold is confirmed by the human expert to be merged, the other nodes, lower in the dendrogram, will not be investigated. Second, a **criteria exploring the tree per branch (*All*)** that is illustrated in Figure 2c. Nodes are still considered according to their ranked *delta* but the dendrogram is explored in more de-

tails. If the human expert confirms a merge on a node ("yes" answer), the lower nodes in the two branches will not be investigated for splitting. If the human experts confirms a separation on a node ("no" answer), the upper nodes will not be investigated for grouping (but can be investigated for splitting). The *2c criteria* relies on a high confidence in the *delta* ranking (the estimation of the distance between x-vectors) and strongly limits the number of questions, while the *All* criteria leads to more questions and thus a finer correction of the dendrogram.

To facilitate the work of the user answering the question, we consider that the HASDS proposes two audio segments (*samples*), for the user to listen to; one for each branch of the current node. Each branch, can link several segments, even for nodes located at the very bottom of the tree (remember that, due to the sequential HAC clustering process, leaves of the dendrogram are clusters linked by the BIC-HAC clustering). The system must select the two most representative or informative *samples*. We investigate 5 *sample* selection methods:

Longest selects the longest segment from each cluster. It assumes that x-vectors from those segments are more robust and that the gain provided by the correction would lead to higher improvement of DER.

Cluster center selects the closest segment to cluster center assuming this is the best representation of this cluster. The center is selected according to the euclidean distance between segments' x-vectors.

Max / Min selects the couple of segments, one from each branch, with the lowest (max) or highest (min) similarity in terms of PLDA score (distance).

Random as a contrastive criteria, a random segment is selected from each cluster (statistics from this method are consolidated by repeating experiments 20 times).

3.3. User simulation and correction module

The correction module simply remembers the successive corrections provided by the human expert. The human expert is simulated for reproducibility and makes use of the ground truth reference to provide a correct answer to each question. To establish a lower bound, we also consider an **ideal** correction method. When a node has been chosen to be investigated, the optimal correction (merging or splitting) is found by looking at the ground truth (reference) to maximize the gain in terms of DER.

4. Experiment: protocol and results

Experiments are performed on the ALLIES dataset[1], an extension of previously existing corpora [31, 32, 33], that includes a collection of 1,008 French TV and Radio shows partitioned in three non-overlapping parts whose statistics are provided in Table 1. The performances are reported as weighed diarization error rate (DER) [34], averaged over all documents of the collection according to their annotated duration. Penalized DER [27] described in Equation 1 is reported as a unique performance indicator including both final DER and human interaction cost.

$$DER_{pen} = \frac{T_{miss} + T_{false} + T_{confusion} + N \cdot t_{pen}}{T_{total}} \quad (1)$$

T_{miss}, T_{false} and $T_{confusion}$ are respectively the duration of missed speech, non-speech considered as speech and wrongly classified speech. T_{total} is the total speech duration in the document. N is the number of corrections applied to the document and t_{pen} is the estimated time a human spend answering one question (see [27]). The quality of the questioning module is estimated by computing the correction/question ratio (CQR) between the number of corrections (question that leads to a modification of the clustering) and the number of questions asked to the human expert. The training set is used to train the x-vector extractor and the PLDA model while the Dev is used to set the clustering threshold. performance are reported on the Eval set.

[1]Database and protocols will be made publicly available after the ALLIES 2021 challenge

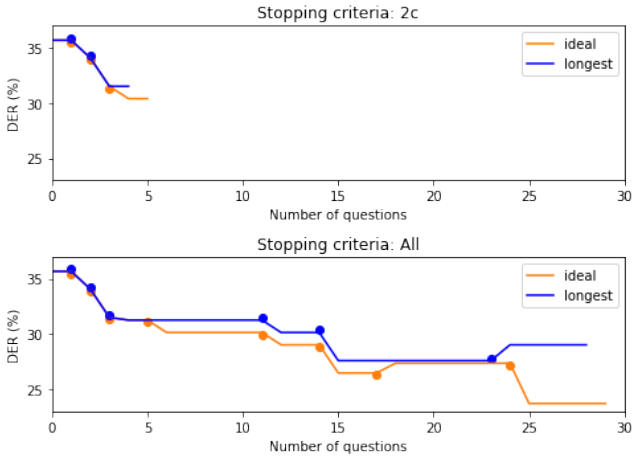


Figure 3: *Tracking of the DER corresponding to a single show file (with duration of 1 hour and 11 minutes) by applying question-correction with different methods. Points in each question indicate that it has resulted in a correction.*

Table 1: *ALLIES dataset description, all duration are given in hh:mm:ss, speakers are considered recurrent when they appear in 3 episodes or more across the dataset.*

Partition	Total Duration	Annotated Duration	#speaker	#recurrent speaker	#shows
Training	223:37:17	175:22:04	3,680	355	475
Dev	105:51:06	33:54:15	983	105	200
Eval	282:36:16	118:53:34	1,720	220	333

4.1. Experiments

Figure 3 illustrates the evolution of DER for an audio file for both stopping criteria. As expected, *All* leads to more questions and achieve a better final DER (lower) than *2c criteria*. This example shows the necessity of taking into account the cost of human interaction to fairly compare HASDS systems.

A first experiment is performed with **ideal** correction to compare the benefit of merging or splitting clusters. Results in Table 2 reveal that for both stopping criteria, DER reduces more when splitting clusters than merging them, but also that the CQR is higher when splitting. Both kind of correction can lead to conflicts. For instance, a node could be merged first while one of its child nodes would requires to be split due to non-purity of the clusters (even with **ideal** correction). For this reason and considering the higher benefits of splitting compared to merging, we chose to prioritize splitting to merging. So if a node has been split into two clusters, its parent nodes will not

be investigated for merging. It helps to avoid investigating the nodes that will not be used for correction.

Table 2: *Performance of the HASDS using **ideal** correction when applying only one type of correction. Second column is the average number of questions per hour and last column reflects is the quality of interaction (CQR).*

Stopping criteria		DER	Avg. #Question / hour	CQR
2c criteria	Merging	15.58	3.75	16.20
	Splitting	11.36	5.94	61.02
All	Merging	15.02	19.86	6.50
	Splitting	10.72	9.49	45.65

A second experiment is performed to compare the 5 *sample* selection methods for both stopping criteria. Results are presented in Table 3. As expected, *ideal* correction provides the lowest DER for both stopping criteria and all 5 proposed methods perform at least as well as the contrastive random selection process. It appears that the longest segments or cluster centers are the most representative from their cluster and that **Min/Max** provide the smaller improvement probably due to the similarity between those criteria and the clustering criteria used by the HAC algorithm. The 5 selection methods are comparable in terms of number of questions asked per hour of audio and CQR. This is visible on the penalized DER which preserves the conclusions drawn by observing the DER.

Comparing the two stopping criteria, we observe that **Longest** and **Cluster center** selection method using the *All* criteria achieves better performance than the *2c criteria* but both criteria achieve similar performance for **Min/Max**. Penalized DER shows that although the *All* criteria achieves lower DER than *2c criteria*, the cost of human interaction (for a t_{pen} empirically set to 6s) for the *All* criteria is much higher and that *2c criteria* might be a better compromise to reduce human interaction. The proposed approach considering *2c criteria* and

Table 3: *DER improvement using different stopping criteria and segment selection methods*

Method	Stopping criteria	DER	Avg. #Q / h	CQR	DER_{pen}
Baseline	-	16.46	-	-	-
Ideal	*2c criteria*	10.57	9.69	43.68%	12.18
	All	9.65	28.21	19.86%	14.35
Random (20 times)	*2c criteria*	12.77±0.13	9.66	44.15%	14.38
	All	12.99±0.16	28.26	22.01%	17.75
Longest	***2c criteria***	**11.18**	9.66	43.56%	**12.79**
	All	**10.45**	28.14	19.97%	15.14
Cluster center	*2c criteria*	11.28	9.64	42.66%	12.89
	All	10.52	28.10	20.17%	15.21
Max	*2c criteria*	12.52	9.69	43.98%	14.13
	All	12.99	28.12	21.40%	17.68
Min	*2c criteria*	12.74	9.61	43.58%	14.35
	All	12.80	28.14	22.33%	17.49

a selection of the longest segment leads to a reduction of DER from 16.46% (without human in loop) to 11.18% while asking less than 10 questions to the human expert perhour of speech processed. However, we found that only 43.56% of the questions asked lead to a correction (i.e., in 56.44% of the cases, the human validates the decision of the automatic system) which will be investigated in future work.

4.2. Analysis

In order to further improve our approach, we analysed the correlation between the benefit of human active correction (in terms of DER reduction or number of questions asked) and the characteristics of the processed audio files (number of speakers, duration of the file...) based on Pearson correlation coefficient, no strong correlation has been found (all less than 0.4).

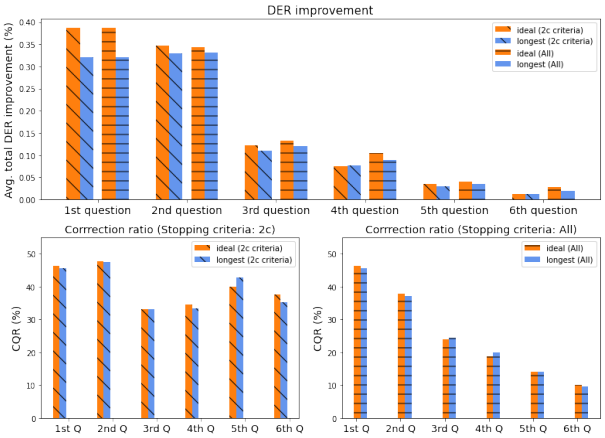


Figure 4: *The performance of questioning based on DER improvement and ratio of correction number to question number.*

We then evaluated the usefulness of successive questions in an ordinal way for each audio file. The total DER improvement and the ratio: number of corrections over the number of questions asked (CQR) are compared base on the question order on Figure 4. For both stopping criteria, the first questions lead to larger DER reductions (upper figure). Interestingly, we can see that the questions asked when using the *2c criteria* have a similar CQR ratio, meaning that successive questions keep contributing to the DER reduction (bottom left). On the other hand, we observe that the CQR reduces for the *All* criteria, meaning that the system tends to ask less useful questions to the expert.

5. Conclusion

The benefit of human active correction for speaker diarization has been investigated. This preliminary study has focused on an active correction of HAC clustering errors. Starting from a strong automatic baseline, we proposed two criteria to ask questions to a human expert. 5 methods to select samples for auditory tests have been proposed and evaluated using a large and challenging dataset that will be publicly released.

Performance of our human assisted speaker diarization system have been evaluated by using a penalized DER proposed in [27] and shows that it can decrease by up to 22,29% relative when applying active correction with the *2c criteria*. This leads to a reduction of 32,07% relative without taking into account the cost of human interaction. The second proposed stopping criteria (*All*) can achieve a relative reduction of 36,51% of DER but requires a higher and less efficient human effort.

This preliminary study is very promising and opens large avenues for future studies. More analyses are ongoing to understand and refine the stopping criteria depending on the nature of the processed audio file and its difficulty for diarization systems. Current studies are conducted to improve the question generation module by estimating the quality of the question before soliciting the human expert. We are also developing the adaptation process in order to improve the automatic system using the information provided by the human expert.

A limitation of this work comes from the restriction to HAC clustering when many works in the literature have been explor-

ing active learning for other clustering algorithms. So far, we have only considered diarization of isolated files but it is very likely that a higher benefit can be expected from active learning when applied to the diarization of a collection of audio files.

6. References

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[2] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Multistage speaker diarization of broadcast news," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1505–1512, 2006.

[3] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote. audio: neural building blocks for speaker diarization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7124–7128.

[4] P.-A. Broux, D. Doukhan, S. Petitrenaud, S. Meignier, and J. Carrive, "Computer-assisted speaker diarization: How to evaluate human corrections," in *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*, 2018.

[5] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The second dihard diarization challenge: Dataset, task, and baselines," *Proc. Interspeech 2019*, pp. 978–982, 2019.

[6] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *Ai Magazine*, vol. 35, no. 4, pp. 105–120, 2014.

[7] L. Jiang, S. Liu, and C. Chen, "Recent research advances on interactive machine learning," *Journal of Visualization*, vol. 22, no. 2, pp. 401–417, 2019.

[8] M. Wang and X.-S. Hua, "Active learning in multimedia annotation and retrieval: A survey," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 2, pp. 1–21, 2011.

[9] C. Yu and J. H. Hansen, "Active learning based constrained clustering for speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2188–2198, 2017.

[10] B. Mateusz, J. Poignant, L. Besacier, and G. Quénot, "Active selection with label propagation for minimizing human effort in speaker annotation of tv shows," in *Workshop on Speech, Language and Audio in Multimedia (SLAM 2014)*.

[11] S. H. Shum, N. Dehak, and J. R. Glass, "Limited labels for unlimited data: Active learning for speaker recognition," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[12] G. Riccardi and D. Hakkani-Tur, "Active learning: Theory and applications to automatic speech recognition," *IEEE transactions on speech and audio processing*, vol. 13, no. 4, pp. 504–511, 2005.

[13] H. Jiaji, C. Rewon, R. Vinay, L. Hairong, S. Sanjeev, and C. Adam, "Active learning for speech recognition: The power of gradients," in *The 30th Conference on Neural Information Processing Systems, NIPS. Barcelona, Spain*, 2016, pp. 1–5.

[14] J. Bang, H. Kim, Y. Yoo, and J.-W. Ha, "Efficient active learning for automatic speech recognition via augmented consistency regularization," *arXiv preprint arXiv:2006.11021*, 2020.

[15] E. Yilmaz, M. McLaren, H. van den Heuvel, and D. A. van Leeuwen, "Language diarization for semi-supervised bilingual acoustic model training," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 91–96.

[16] D. G. Karakos, S. Novotney, L. Z. 0002, and R. M. Schwartz, "Model adaptation and active learning in the bbn speech activity detection system for the darpa rats program." in *INTERSPEECH*, 2016, pp. 3678–3682.

[17] M. Abdelwahab and C. Busso, "Active learning for speech emotion recognition using deep neural network," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 1–7.

[18] B. Settles, "Active learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2009.

[19] M. Wang, F. Min, Z.-H. Zhang, and Y.-X. Wu, "Active learning through density clustering," *Expert systems with applications*, vol. 85, pp. 305–317, 2017.

[20] S. Basu, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering," in *Proceedings of the 2004 SIAM international conference on data mining*. SIAM, 2004, pp. 333–344.

[21] P. K. Mallapragada, R. Jin, and A. K. Jain, "Active query selection for semi-supervised clustering," in *2008 19th International Conference on Pattern Recognition*. IEEE, 2008, pp. 1–4.

[22] Q. Xu, K. L. Wagstaff *et al.*, "Active constrained clustering by examining spectral eigenvectors," in *International Conference on Discovery Science*. Springer, 2005, pp. 294–307.

[23] X. Wang and I. Davidson, "Active spectral clustering," in *2010 IEEE International Conference on Data Mining*. IEEE, 2010, pp. 561–568.

[24] S. Miyamoto and A. Terami, "Semi-supervised agglomerative hierarchical clustering algorithms with pairwise constraints," in *International Conference on Fuzzy Systems*. IEEE, 2010, pp. 1–6.

[25] I. Davidson and S. Ravi, "Agglomerative hierarchical clustering with constraints: Theoretical and empirical results," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2005, pp. 59–70.

[26] E. Geoffrois, "Evaluating interactive system adaptation," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 256–260.

[27] Y. Prokopalo, S. Meignier, O. Galibert, L. Barrault, and A. Larcher, "Evaluation of lifelong learning systems," in *International Conference on Language Resources and Evaluation*, 2020.

[28] P.-A. Broux, F. Desnous, A. Larcher, S. Petitrenaud, J. Carrive, and S. Meignier, "S4d: Speaker diarization toolkit in python," 2018.

[29] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Phonetically-constrained plda modeling for text-dependent speaker verification with multiple short utterances," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7673–7677.

[30] A. Larcher, A. Mehrish, M. Tahon, S. Meignier, J. Carrive, D. Doukhan, O. Galibert, and N. Evans, "Speaker embeddings for diarization of broadcast data in the allies challenge," in *submitted to 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

[31] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The repere corpus: a multimodal corpus for person recognition." in *LREC*, 2012, pp. 1102–1107.

[32] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ester phase ii evaluation campaign for the rich transcription of french broadcast news," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[33] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert, "The etape corpus for the evaluation of speech-based tv content processing in the french language," in *International Conference on Language Resources, Evaluation and Corpora*, 2012.

[34] O. Galibert, "Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech." in *INTERSPEECH*, 2013, pp. 1131–1134.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

An Automatic System for Dementia Detection using Acoustic and Linguistic Features

Miriam Gonzalez-Atienza, Jose A. Gonzalez-Lopez, and Antonio M. Peinado

Dept. of Signal Theory, Telematics and Communications
University of Granada

myriamgonzalez@correo.ugr.es, joseangl@ugr.es, amp@ugr.es

Abstract

Early diagnosis of dementia is crucial for mitigating the consequences of this disease in patients. Previous studies have demonstrated that it is possible to detect the symptoms of dementia, in some cases even years before the onset of the disease, by detecting neurodegeneration-associated characteristics in a person's speech. This paper presents an automatic method for detecting dementia caused by Alzheimer's disease (AD) through a wide range of acoustic and linguistic features extracted from the person's speech. Two well-known databases containing speech for patients with AD and healthy controls are used to this end: DementiaBank and ADReSS. The experimental results show that our system is able to achieve state-of-the-art performance on both databases. Furthermore, our results also show that the linguistic features extracted from the speech transcription are significantly better for detecting dementia.
Index Terms: Acoustic voice analysis, speech-based disease diagnosis, dementia, Alzheimer's disease, word embeddings.

1. Introduction

Dementia is a type of neurodegenerative disease, whose most common cause is Alzheimer's Disease (AD). Although memory impairment is the main early symptom for AD, it has been found that language and speech abilities also decline, even in the very early stages of the disease [1]. This is known to affect object naming, noun production and rates of verb usage. In general, loss of vocabulary, simplified syntax/semantics, and overuse of semantically empty words are commonly found in the language of people with dementia [2, 3]. Speech is, therefore, a promising candidate as a source of information for new approaches to diagnosing dementia. As no curative therapy is known, early secondary prevention measures are of great importance. Examinations typically include a large number of neuropsychological tests, which are very time consuming and expensive. In order to enable longitudinal cognitive status monitoring on a large scale, a fast and cheap method for diagnosing the disease needs to be found.

Automatic speech processing has been shown to be a promising way in the diagnosis of dementia. Approaches have used acoustic, prosodic and linguistic features [4, 5, 6] in a classification task that aims to distinguish people affected by dementia from cognitively healthy subjects using just their speech. Patient's speech and language are obtained from written texts and speech recordings. For example, *picture description* is a constrained task that relies less on episodic memory, but requires more semantic knowledge and retrieval ability. The most commonly used picture prompt is a line drawing called "Cookie Theft" [7]. During the test, the patients are asked to describe what they see in the picture, while the answer is recorded. As reported [8], people suffering from dementia tend to hesitate more often and make longer pauses. Thus, features based on the occurrence and duration of pauses, extracted from the output of a Voice Activity Detector (VAD), are very promising for automatic detection of this disease.

In this paper, we present the details of our system for dementia detection using features extracted from speech recordings and its corresponding transcriptions. We use both acoustic and linguistic features and compare the results to those obtained in previous investigations. Feature selection of the most relevant features is performed in order to achieve better classification results. We evaluated our proposed system on two well-known databases containing speech material recorded by AD patients and healthy controls (HC): DementiaBank [9] and the recently released Alzheimer's dementia recognition through spontaneous speech (ADReSS) database [10].

This paper has the following structure. In Section 2, we present the details of our system for automatic dementia detection. Section 3 is dedicated to explain the process followed to select the most important features in classification. The experimental results are presented in Section 5, whereas the main conclusions of this work are listed in Section 7.

2. Dementia detection system

Figure 1 shows a block diagram for the proposed automatic system for speech-based dementia detection. Firstly, the voice of the participant is recorded and transcribed (manually, in our case) while the subject performs a cognitive task (i.e., describing a drawing such as the Cookie Theft from the Boston test). Then, a noise reduction technique is applied to the audio signal. The enhanced signal is passed to the feature extraction block where a set of acoustic and linguistic features are extracted. Finally, a selection of this features are sent to a machine learning classifier trained to discriminate between AD and HC subjects.

The details of our system are given in the following sections. Table 1 shows a summary of the acoustic and linguistic features extracted by our system for automatic dementia detection.

2.1. Acoustic features

Acoustic features measure how participants speak. They are extracted from the de-noised speech signals after applying the spectral noise gating method described in [11] to the audio signals to improve their quality. In our system, we consider the

This work was funded by the Spanish State Research Agency (SRA) under the grant PID2019-108040RB-C22/SRA/10.13039/501100011033. Jose A. Gonzalez-Lopez holds a Juan de la Cierva-Incorporation Fellowship from the Spanish Ministry of Science, Innovation and Universities (IJCI-2017-32926).

10.21437/IberSPEECH.2021-56

Table 1: *Description of the 319 speech features extracted by our system for automatic dementia detection.*

Feature set	Description	No.
Pause based	% of pauses in speech	1
	% of pauses in utterance	1
	Speech and pauses duration	2
Speech rhythm	Beats per minute	1
Spectral features	MFCCs mean	1
	MFCCs variance	1
	MFCCs skewness	1
	MFCCs kurtosis	1
	Mean of ΔMFCCs	1
	Variance of ΔMFCCs	1
	Mean of first 24 spectral centroids	1
Prosodic features	Avg. of F_0 values	1
	Std. of F_0 values	1
	Acoustic features:	**14**
Word count	% of unique words	1
POS tags	% of verbs	1
	% of determinants	1
	% of nouns	1
Word embeddings	Avg. of word embeddings	300
	Std. of word embeddings	1
	Linguistic features:	**305**

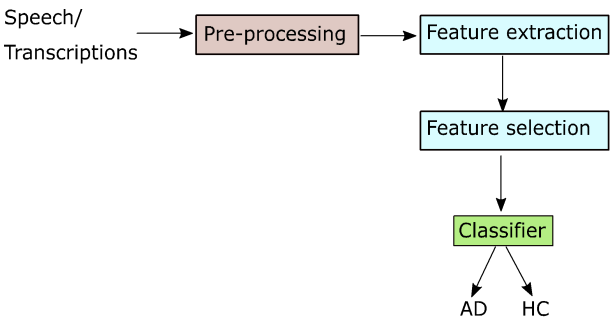


Figure 1: *Block diagram of the dementia detection system*

following four types of acoustic features:

Pause-based features: People afflicted with dementia tend to hesitate more often and make longer pauses than HC subjects [6, 12]. Thus, we computed four pause-based features for each utterance, including total pause and speech duration (in seconds), rate between total duration of pauses and duration of the utterance, and rate between the total duration of pauses and the duration of speech for each utterance. These features were automatically computed from the outputs of an energy-based VAD technique applied to the de-noised signals. The energy threshold was manually defined as 60% of the signal energy.

Speech rhythm: Longer pauses and more hesitations in speech from people with dementia imply that more time is needed to convey words. Thus, we measured the speech rhythm as number of beats per minute using the Predominant Local Pulse (PLP) algorithm, first introduced by Grosche and Müller [13].

Spectral features: Mel-frequency cepstral coefficients (MFCCs) [14, 11, 15, 16] were extracted from windows of 25 ms with 15 ms overlap to capture the spectral content of the speech signal. We then computed the average across time for the first 24 MFCCs. Then, we computed the mean, variance, skewness and kurtosis of the resulting average vector. We also extracted their first order derivatives with their mean and variance. Spectral centroid parameters[17], which aim to locate the spectrum centre of mass and have been found to be valuable in measuring the cognitive load [18, 11], were also extracted from the audio waveforms and computed their mean.

Prosodic features: We computed the speech fundamental frequency (F_0) using the autocorrelation method [19]. From the F_0 values, we computed the mean and standard deviation through the signal.

2.2. Linguistic features

Linguistic features are used to measure changes in vocabulary and sentence structure that are caused by dementia. Our linguistic features operate at the word level of transcriptions. In this work, the manual transcriptions provided with the databases are used to extract the linguistic features. In general, AD patients tend to make shorter phrases than HC and also have a less-rich vocabulary. The linguistic features described below are computed as proportions considering the total number of words spoken in each utterance, e.g., percentage of adjectives w.r.t. the total number of spoken words.

Word count: Word count is a common used feature to classify people with dementia [16]. Verbal repetition is a hallmark of dementia and AD at all stages, but is most commonly targeted for monitoring and treatment effects in its mild stage [20]. For that reason, we have extracted the proportion between unique words and the total number of words in each transcription.

Part-of-Speech (POS) Tags: Words with similar grammatical properties can be grouped together by POS tags. Each tag represents the grammatical role a word can take in a sentence and thus POS tags can be used to indicate grammatical properties of participant's speech. We used TreeTagger [21] to automatically extract POS tags and calculate the frequency of occurrence of each tag. We have considered three POS categories: verbs, determinants and nouns. Once the frequency of occurrence is calculate for each participant, we measure the proportion between the number of POS categories and the number of words in each sentence.

Word embeddings: Word embedding is a technique widely used to convert written words into feature vectors. Recently, successful approaches have used deep learning techniques to produce vectors representing words. In this work, we have used the *word2vec* technique [22], which is based on the co-ocurrences of words taking into account the context, to extract 300-dimensional embeddings for every word spoken by the participant in each sentence. Then, we average all the embeddings obtained in each sentence to finally obtain a 300-dimensional vector representing the linguistic content of that sentence. Besides, we calculate the mean and standard deviation for each word embedding vector of the subject. Thus, we obtain a 300-dimensional vector for each subject, with two additional values: the mean and standard deviation of this vector.

3. Feature selection

We applied a feature selection procedure to select the most meaningful acoustic and linguistic features and discard the features that contribute less to the classification accuracy. We trained an extremely randomized trees (ERT) classifier [23], which is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result. Feature selection was applied to the training set and then, the optimum

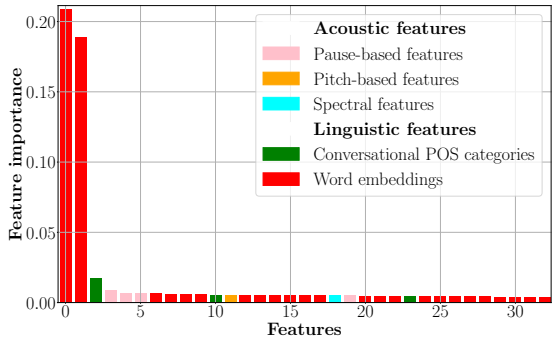


Figure 2: *Ranking of the top 15% features for DementiaBank database.*

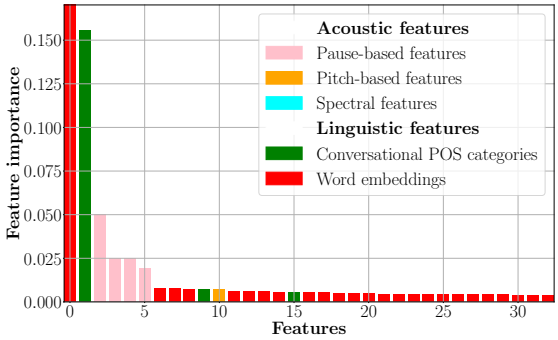


Figure 3: *Ranking of the top 15% features for the ADReSS database.*

set of features selected by this procedure, was used during evaluation time to detect dementia based on a person's speech.

Figures 2 and 3 show the ranking of the top 15% features selected by the the ERT technique for the DementiaBank and ADReSS databases, respectively, from the whole set of 319 acoustic and linguistic features extracted by our system. As can be seen, the majority of selected features are linguistic. This, as also reported by other authors [16, 20], makes sense because dementia is known to have a profound impact on a person's speech, even from its first stages [1]. In particular, POS features based on the proportions of nouns, verbs, and determinants are very relevant for detecting dementia on both databases, with relative importance of 0.01, 0,005 and 0.004 in DementiaBank, and 0.150, 0.015, and 0.014 in ADReSS. Word embedding based features are also among the best features for both databases, thus highlighting again the importance of linguistic-based features for the task of automatic detection of neurodegenerative diseases. Although acoustic features are deemed less relevant, the ERT technique also selected a significant number of them, particularly for the ADReSS database, where pause-based features related to pause and speech ratio are among the top 5-best features.

4. Experimental setup

4.1. Databases

To evaluate the proposed dementia detection system, we used the audio recordings and transcriptions from the DementiaBank [11, 24] and ADReSS [10] databases. DementiaBank is a free-access, large existing database for Alzheimer's and related dementia diseases collected during longitudinal study conducted by the University of Pittsburgh School of Medicine. Verbal descriptions of the Boston Cookie Theft picture were recorded from people with different types of dementia with an age span from 49 to 90 years as well as from elderly healthy control subjects within an age range from 46 to 81 years. During the interviews, patients were given the picture and were told to discuss everything they could see happening in the picture. There are a total of 473 recordings from 97 healthy controls and 233 speech samples from 167 AD patients diagnosed as possible or probable AD. We splitted this database randomly into training and evaluation subsets. In the training subset, we have selected 150 subjects, as well as for the evaluation subset, in order to have the same number of subject in both groups.

Similarly, the ADReSS database is a subset of DementiaBank with acoustically enhanced audio recordings and matched in terms of age and gender (i.e., major factors in recognising dementia) to avoid bias towards them. The dataset contains speech recordings from 78 non-AD subjects and 78 AD subjects while describing the Cookie Theft drawing. This database already defines a training subset, containing data for 54 AD and 54 HC subjects, and an evaluation subset with 24 subjects for each group.

A summary of the characteristics of both databases in terms of number of subjects in each class in shown in Table 2.

Table 2: *Basic characteristics of the patients in each group in the ADReSS challenge dataset and DementiaBank*

Dataset		**No. subjects**	
		AD	**Non-AD**
ADReSS	Train	54	54
	Test	24	24
DementiaBank	-	322	229

4.2. Evaluation

We evaluated four state-of-the-art classifiers for the task of classifying dementia from the set of extracted linguistic and acoustic features: Linear Discriminant Analysis (LDA), Support Vector Machines (SVMs), Random Forests and Adaptive Boosting, previously used for dementia detection in [12, 25, 26]. We also evaluated the effect of the feature selection procedure described in Section 3 on classification accuracy. In particular, the following configurations were evaluated: (i) using only either acoustic features (Acoustic) or (ii) linguistic features (Linguistic) with no feature selection; or (iii) a combination of both types of features where only a fraction of them are used for classification, as provided by the feature selection algorithm (Feature selection).

The following classification metrics are reported for each configuration: accuracy, recall (sensitivity) and specificity, also considered in [1, 11, 27]. Accuracy is the percentage of correct predictions. Sensitivity describes what proportion of patients with AD are correctly identified as having AD, while specificity describes what proportion of HC persons are correctly identified as belonging to that class.

5. Experimental Results

Table 3 shows the classification results achieved by our system on the DementiaBank dataset. With LDA classifier and using only linguistic features, we achieve an accuracy of 96%, 97% sensitivity and 92.3% specificity, providing the higher classifi-

Table 3: *Classification results for DementiaBank dataset. Accuracy/Sensitivity/Specificity (%). In bold are shown the higher classification results by rows*

Features	LDA	SVM	RF	AdaBoost	Avg.
Acoustic	65/65/78	67/65/60	**71/71/70**	65/65/73	67/67/70
Linguistic (with word embeddings)	**96/97/92**	62/62/92	74/60/93	78/78/75	78/74/88
Feature selection (15%)	**95/95/93**	67/65/57	78/56/48	81/77/74	80/73/68
Feature selection (25%)	**93/93/94**	75/74/81	81/67/64	78/78/74	82/78/78
Feature selection (50%)	**87/86/94**	83/83/88	78/66/68	68/67/75	79/76/81
Feature selection (75%)	**94/94/95**	72/71/71	81/67/64	85/78/77	83/78/77
Feature selection (100%)	82/82/71	71/65/86	**90/87/85**	79/66/62	81/75/76
Al-Hammed et al [11, 14]	-	86/-/-	96/-/-	86/-/-	-

Table 4: *Classification results for ADReSS dataset. Accuracy/Sensitivity/Specificity (%). In bold are shown the higher classification results by rows*

Features	LDA	SVM	RF	AdaBoost	Avg.
Acoustic	65/65/67	**65/65/78**	59/57/78	59/58/78	62/61/75
Linguistic (with word embeddings)	72/71/80	**78/76/67**	61/59/80	65/65/78	69/68/76
Feature selection (15%)	76/55/54	**79/65/60**	57/56/80	57/55/80	67/58/69
Feature selection (25%)	53/53/50	**66//60/56**	57/55/80	56/54/75	58/56/65
Feature selection (50%)	59/53/47	55/53/50	**69/68/80**	57/58/80	60/58/64
Feature selection (75%)	**65/58/54**	61/58/56	58/58/80	56/57/80	60/58/68
Feature selection (100%)	60/55/50	62/60/58	**66/65/80**	57/53/80	62/58/67
Martinc *et al.* [28]	77/-/-	51/-/-	55/-/-	-	
Luz *et al.* [10]	-	75/-/-	-	-	-

cation metrics among all the classifiers and features combinations. Furthermore, the best classification results with Feature Selection are obtained when using only 15% of the features. Although we achieve high classification metrics with Feature Selection, the best classification results are obtained for the configuration using the LDA classifier and linguistic features only.

Table 4 shows the classification results obtained on the ADReSS dataset. With SVM classifier and 15% of the features, we achieve a maximum accuracy of 79%, 65% sensitivity and 60% specificity. As well as with DementiaBank dataset, the best classification results are obtained when a small subset of the features is selected. This is also the case for related works, such as Luz *et al.*[10], where the accuracy for their best performing system drops 4% (relative) when feature selection is not performed on their original set of 370 features. Again, linguistic features showed more capability to differentiate between HC and AD than the acoustic features.

6. Discussion

From the comparison between acoustic and linguistic features, we conclude that linguistic features provide significantly better classification results. Hence, they present more capability to distinguish between HC and AD than acoustic features. Recently and similar to our work, Fraser *et al.* [16] studied the potential of using linguistic features to identify Alzheimer's disease. In total, a set of 370 acoustic, lexical and semantic features were extracted and they obtained a highest accuracy of 92% in distinguishing between HC subjects and AD patients using the top 25 ranked features.

From the comparison between the results obtained on the ADReSS and DementiaBank datasets, it can be observed that better classification results are obtained on DementiaBank. One possible explanation is the number of training subjects in each dataset, which is considerably larger in DementiaBank (300 subject in DementiBank vs. 108 in ADReSS). Another possible explanation could be that ADReSS is gender and age balanced, whereas DementiaBank is not. Thus, it could be that, in the case of DementiaBank, the classifiers are able to infer the age and gender of each subject from the e.g., acoustic features to achieve a better performance. Besides, the recordings from DementiaBank have a higher level of background noise than the recordings from ADReSS, which causes that acoustic features contribute less to the classification results in comparison with the acoustic features on ADReSS dataset. Finally, as can be seen from Table 3 and 4, we have achieved higher classification results than the ones presented in the literature.

7. Conclusions

In this paper, we have proposed a binary classifier approach based on speech for dementia detection using acoustic and linguistic features derived from audio recordings and text transcriptions. Feature Selection has showed that the best features for this task are pause-based features, word embeddings and Conversational POS categories (nouns, verbs and determinants). We have evaluated our system on two well-known databases achieving state-of-the-art results for both. We have obtained the best classification results with LDA classifier and linguistic features for DementiaBank dataset, achieving an accuracy of 96%, a sensitivity of 97% and 92% of specificity. On the other hand, with ADReSS dataset we have achieved the maximum classification results with SVM classifier and 15% of the features, with an accuracy of 79%, a sensitivity of 65% and 60% of specificity. We therefore conclude from this results that linguistic features have the capability to distinguish from people with AD and healthy control.

In the future, we plan to investigate the use of more features as well as other machine learning algorithms. Furthermore, it would be interestint to evaluate the robustness of the system for detecting dementia in other databases or, even, in different languages.

8. References

[1] S. Ahmed, A.-M. F. Haigh, C. A. de Jager, and P. Garrard, "Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease," *Brain*, vol. 136, no. 12, pp. 3727–3737, 2013.

[2] J. Appell, A. Kertesz, and M. Fisman, "A study of language functioning in alzheimer patients," *Brain and Language*, vol. 17, no. 1, pp. 73–91, 1982.

[3] R. Jaffard, "Communication and cognition in normal aging and dementia," vol. 28, pp. 229–230, 1990.

[4] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance," vol. 14, pp. 71–91, 2000.

[5] A. Khodabakhsh, F. Yesil, E. Guner, and C. Demiroglu, "Evaluation of linguistic and prosodic features for detection of alzheimer's disease in turkish conversational speech," vol. 2015, 2015, j AUDIO SPEECH MUSIC PROC. 2015, 9 (2015).

[6] J. Weiner and T. Schultz, *Selecting Features for Automatic Screening for Dementia Based on Speech: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings*, 01 2018, pp. 747–756.

[7] L. Cummings, "Describing the cookie theft picture: Sources of breakdown in alzheimer's dementia," *Pragmatics and Society*, vol. 10, no. 2, pp. 153–176, 2019.

[8] J. Rodríguez, H. Martínez, and B. Valles, "Las pausas en el discurso de individuos con demencia tipo alzheimer. estudio de casos," *Revista de investigación en logopedia*, vol. 5, no. 1, pp. 40–59, 2015.

[9] [Online]. Available: https://talkbank.org/DementiaBank/. [Accessed:10-Dec- 2015]Dementia Bank."

[10] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge," in *Proc. Interspeech 2020*, 2020, pp. 2172–2176. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2571

[11] S. Al-Hameed, M. Benaissa, and H. Christensen, "Simple and robust audio-based detection of biomarkers for alzheimer's disease," sLPAT 2016 Workshop on Speech and Language Processing for Assistive Technologies. 13 September 2016, San Francisco, USA.

[12] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of alzheimer's disease in conversational german," in *Proc. Interspeech 2016*, 2016, pp. 1938–1942. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-100

[13] P. Grosche and M. Müller, "Extracting predominant local pulse information from music recordings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, pp. 1688 – 1701, 09 2011.

[14] S. Al-Hameed, M. Benaissa, and H. Christensen, "Detecting and predicting alzheimer's disease severity in longitudinal acoustic data," in *Proceedings of the International Conference on Bioinformatics Research and Applications 2017*, ser. ICBRA 2017. New York, NY, USA: Association for Computing Machinery, 2017, p. 57–61. [Online]. Available: https://doi.org/10.1145/3175587.3175589

[15] S. Luz, "Longitudinal monitoring and detection of alzheimer's type dementia from spontaneous speech data," in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2017, pp. 45–46.

[16] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech." *Journal of Alzheimer's disease : JAD*, vol. 49, pp. 407–422, 2016.

[17] P. N. Le, E. Ambikairajah, J. Epps, V. Sethu, and E. H. Choi, "Investigation of spectral centroid features for cognitive load classification," *Speech Communication*, vol. 53, no. 4, pp. 540–551, 2011.

[18] S. Al-Hameed, M. Benaissa, H. Christensen, B. Mirheidari, D. Blackburn, and M. Reuber, "A new diagnostic approach for the identification of patients with neurodegenerative cognitive complaints," vol. 14, p. e0217388, pLoS One. 2019;14(5):e0217388. Published 2019 May 24.

[19] L. Tan and M. Karnjanadecha, "Pitch detection algorithm: autocorrelation method and amdf," in *Proceedings of the 3rd international symposium on communications and information technology*, vol. 2, 2003, pp. 551–556.

[20] E. Reeve, P. Molin, A. Hui, and K. Rockwood, "Exploration of verbal repetition in people with dementia using an online symptom-tracking tool," vol. 29, pp. 959–966, 2017.

[21] H. Schmid, "Improvements in part-of-speech tagging with an application to german," pp. 13–25, 1999.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, pp. 3111–3119, 2013.

[23] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, p. 3–42, Apr. 2006. [Online]. Available: https://doi.org/10.1007/s10994-006-6226-1

[24] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Detecting signs of dementia using word vector representations," 09 2018, pp. 1893–1897.

[25] M. Yancheva and F. Rudzicz, "Vector-space topic models for detecting Alzheimer's disease," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 2337–2346. [Online]. Available: https://www.aclweb.org/anthology/P16-1221

[26] B. Mirheidari, D. Blackburn, M. Reuber, T. Walker, and H. Christensen, "Diagnosing people with dementia using automatic conversation analysis," 09 2016, pp. 1220–1224.

[27] L. Breiman, "Random forests," vol. 45, pp. 261–277, 2001, machine Learning volume 45, pages5–32(2001).

[28] M. Martinc and S. Pollak, "Tackling the adress challenge: a multimodal approach to the automated recognition of alzheimer's dementia," *Proc. Interspeech 2020*, pp. 2157–2161, 2020.

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Alzheimer's Dementia Detection from Audio and Language Modalities in Spontaneous Speech

Edward L. Campbell[1], *Laura Docío-Fernández*[1], *Javier Jiménez-Raboso*[2], *Carmen García-Mateo*[1]

[1]GTM research group, AtlanTTic Research Center, University of Vigo, Spain
[2]acceXible

ecampbell@gts.uvigo.es, ldocio@gts.uvigo.es, javij@accexible.com, carmen.garcia@uvigo.es

Abstract

Automatic detection of Alzheimer's dementia (AD) by speech processing is enhanced when features of both the acoustic waveform and the content are extracted. Audio and text transcription have been widely used in health-related tasks, as spectral and prosodic speech features, as well as semantic and linguistic content, convey information about various diseases. Hence, this paper describes and compares the performance of different Alzheimer's disease detection approaches based on both the patient's voice and message transcription. To this effect, five different individual systems are analysed: three of them are speech-based and the other two systems are text-based. Specifically, as speech-based systems the x-vector and i-vector paradigm to characterise speech, and a set of rhythmic-based hand-crafted features are proposed. And, for transcription analysis, two systems are proposed, one which uses pre-trained BERT models and the other which uses knowledge-based linguistic and language modelling features. Also, to examine if acoustic and content features are complementary intra-modality and inter-modality score fusion strategies are studied. Experiments in the framework of Interspeech 2020 ADReSS challenge show that the BERT-based system outperforms other individual systems for the AD detection task. Furthermore, the fusion of acoustic- and transcription-based systems provides the best result, suggesting that the two modalities are complementary to some extent.

Index Terms: Alzheimer's disease detection, i-vector, x-vector, speech fluency, BERT, score level fusion, ADReSS challenge

1. Introduction

The Alzheimer's disease (AD) is a neurodegenerative illness that represents the most common cause of dementia in the world. It is provoked by the damage of neurons involved in thinking, learning and memory. This disease has three main stages. In the first one, called preclinical, patients do not present clear symptoms because the brain initially compensate for them, enabling individuals to continue to function normally. The second one is defined as Mild Cognitive Impairment (MCI). In this stage, patients show greater cognitive decline then expected for their age, having problems to express and connect ideas. However, these changes may be only noticeable to family members and friends. The critical phase is the last one, which is called as dementia. It is characterized by noticeable memory, thinking and behavioural symptoms that impair a person's ability to function in daily life [1][2].

Common signs of AD are related to problems with uttering words; consequently, people with AD may have trouble following or joining a conversation, they may stop in the middle of a sentence and have no idea how to continue. As a result, analysis of speech and its transcription may represent a suitable mechanism for detecting the AD during the second or third stage of the disease [1] [3]. According to the literature[4][5][6], using information from the patient's voice, as well as from its transcription would ease the early AD detection task.

Different multimodal approaches for AD recognition have been proposed in the ADReSS challenge [7]. In them, speech is commonly represented using x-vectors, a large set of functionals obtained from low level descriptors, and other deep learning based speech representations. Using the speech transcriptions, best systems [8] are those based on deep language embeddings such as Bidirectional Encoder Representations from Transformers (BERT) [9].

In this work, three speech-based systems and two text-based systems are proposed for automatic distinction between individuals with and without AD. Those based on the speech signal are compound by four approaches to represent their spectral and prosodic content, namely: i-vector and x-vector embeddings, and rhythmic features. The first two use support vector machine (SVM) as classifiers and the last one uses a linear discriminant analysis (LDA) classifier. As for systems that use speech transcriptions, one is based on fine-tuning BERT model [9] for text classification, and the other one is based on features extracted by language modelling, using a SVM as classifier. Finally, an intra-modality and inter-modality score fusion strategy was done to improve the final results. All the individual systems, and their fusion, are evaluated on the AD recognition task within the framework of the ADReSS Challenge [7]. This challenge targets the AD detection using spontaneous speech. The data used in the Challenge consists of speech recordings, and their transcripts, corresponding to the description of the Cookie Theft picture. Specifically, it is a selection of Alzheimer and control patients from of the DementiaBank's Pitt corpus [1].

The rest of the paper is organized as follows. Sections 2 and 3 describe the speech-based and text-based systems, respectively. Section 4 outlines the experimental framework. The experimental results are exposed and discussed in Section 5. Finally, Section 6 draws some conclusions and future work.

2. Speech-based systems

In this section, a description of speech feature extraction strategies and classification methods is done, with special attention to different statistical features extracted from the patient's voice.

2.1. Speech embedding features

Speech embedding features are considered as state-of-art speech representation for speaker recognition application. These speech representations can be applied for AD detection,

[1]http://dementia.talkbank.org/

10.21437/IberSPEECH.2021-57

as long as they preserve those spectral patterns in the speaker's voice that allow the distinction between individuals with and without AD. In this paper, two strategies were analyzed, the first one uses the i-vector paradigm [10], and the second one uses an x-vector [11] based representation. The main characteristics of these approaches are briefly described below.

A. *i-vectors*

To extract the i-vectors a universal background model (UBM) and a total variability matrix $\mathbf{T}$ model must be trained. As speech parameterization these models use 13 perceptual linear prediction (PLP) cepstral coefficients, combined with two pitch-related features (F0 and voicing probability) [12]. This features are augmented with their delta and acceleration coefficients, leading to vectors of dimension 45. These features were chosen since that combination achieves a representation of speech that includes spectral information and prosodic features such as rhythm or intonation that are embedded in the fundamental frequency. The UBM has a diagonal covariance, 512-component gaussian mixture model (GMM), and was trained with data from outside this task; $\mathbf{T}$ was trained using the task training data. The dimension of the i-vectors was set to 125 and they were length-normalized.

B. *x-vectors*

A pretrained time delay deep neural network (TDNN) [2] with 5 time delay layers and two dense layers was used. The network was trained to discriminate between speakers, using the nnet3 neural network library of the Kaldi Speech Recognition Toolkit [13] on augmented VoxCeleb 1 and VoxCeleb 2 datasets [14]. The input to the TDNN are 30 mel-frequency cepstral coefficients (MFCC), and the embeddings are extracted from the first dense layer with a dimensionality of 512. The output of this layer (x-vector) is first projected using latent discriminant analysis (LDA) into a 200 dimensional space and then length-normalized.

Instead of extracting an i-vector or x-vector (embeddings) to represent the entire audio signal, a set of these vectors is obtained applying a sliding window. In this way, each audio file is represented by a certain number of embeddings, which are then used for classification. The optimal window length and overlap were tuned experimentally.

Both systems use as classifier an SVM with a linear kernel. Since a number of embeddings are extracted from each audio, there will also be a set of classification results (one for each embedding), which must be combined to obtain a patient's classification in AD or non-AD. In this work, the mean of the classifications was used as score for the final decision.

2.2. Speech fluency

The lack of speaking fluency is a common pattern in patient with AD, being the rhythm a viable clue to detect that behavior. However, prosodic information (e.g., mean energy) was also used because the correct pronunciation of words does not only depends on the rhythm but also on intonation, tone and stress. For this system, the selected parameters were based on [4] [15], and they are as follows:

- Number of syllables
- Rate of speech (syllables / original duration)
- Speaking duration
- Average fundamental frequency
- Median of fundamental frequency
- Minimum fundamental frequency
- Pronunciation posterior probability
- Average voice interval duration
- Average duration of pairs [3]
- Mean energy
- The ratio between the energy mean and its standard-deviation

The extraction process was done using the Python library My-Voice Analysis [4], the voice activity detection of the SIDEKIT software [16], and our own algorithms.

The classification process was done by the LDA algorithm, projecting the rhythmic feature vector into an one-dimensional space where every projected point represents a classification score.

3. Text-based systems

Two different text-based approaches were analysed. The first approach uses a pre-trained Bidirectional Encoder Representation from Transformers (BERT) sequence classification model [9]. The second approach manually extract linguistic information for creating input features for a classifier.

3.1. Text preprocessing

The transcripts contained in the dataset are in CHAT format [17], which facilitates speech annotation and analysis. They include the transcript of both the subject and the investigator in charge of the test, as well as additional non-speech annotations such as times, pauses, errors, morphological or syntactic analysis. This information is represented with special characters (such as "[//]" for pauses), and since they are very specific for this format they probably are not present in the BERT original tokenizer. We included these tokens in the BERT tokenizer, so a representation of them can be learned during fine-tuning.

Transcripts are divided into several interventions, i.e. sentences or parts of complete sentences with meaning, and this granularity has been maintained in the preprocessing. Metadata, researcher interventions and linguistic analysis included in the files have been removed. We keep both subject's words and some annotations from the transcription (for pauses, errors and subject's actions such as laughing) as input for the classification.

3.2. BERT model

The first approach consists on fine-tuning a pre-trained version of BERT at intervention level, by classifying if a given sentence belongs to an AD subject. By using the interventions as independent samples the training set is increased to a size of 1492 records from the 108 subjects. Then, the probability that the subject had AD given all his/her interventions is given by

$$p(\text{AD}) = \frac{\sum_i l_i s_i}{\sum_i l_i}, \tag{1}$$

where l_i is the length in tokens of the i-th intervention of the subject and s_i is its score (between 0 and 1) estimated by the

[2] http://kaldi-asr.org/models.html

[3] Pairs: consecutive voiced and unvoiced segments

[4] https://github.com/Shahabks/my-voice-analysis

model. With this weighted mean, more importance is given to longer interventions.

Input sequences are tokenized and padded to a maximum length of 40. We used the uncased version of BERT [18][5] for automated feature extraction. The resulting vector of dimension 768 is mapped to a final linear layer to perform the binary classification. A dropout rate of 0.3 is added to the last layer to prevent overfitting. The whole model is trained for 3 epochs, using a batch size of 16 and AdamW optimizer [19] with an initial learning rate of 5e-5 and linear scheduling.

3.3. Model based on linguistic features

In this approach, several linguistic features and indicators are built from subject's interventions and, using this feature vector as input, an SVM is trained. Unlike the previous method, the classification here is performed at subject level, taking the full transcript of each participant, and only subject's words are considered.

Previous works [20] have shown that certain linguistic features are useful for detecting AD using Cookie Theft test. Here 13 features have been built, grouped into 4 categories:

- Extension information such as the number of interventions, number of words per intervention and mean word length.
- Vocabulary richness, by measuring the number of unique words used by the subject.
- Presence of key informational concepts: kitchen, mother, stool, boy and girl.
- Frequency of verbs, nouns, adjectives and pronouns from POS-tagging.

Each feature is then rescaled with min-max normalization in range [0, 1] and an SVM with radial basis function (RBF) kernel and C = 1.0 is trained, whose output is the probability that the subject had AD given the 13-dimensional feature vector.

4. Experimental framework

The training dataset [7] consists of the recordings and manual transcripts of 108 subjects performing the test known as Cookie Theft, whose objective is to describe an image. Out of the 108 participants, 54 are patients diagnosed with Alzheimer's. Table 1 shows the training data distribution in detail.

Table 1: *ADReSS training dataset*

	AD		**non-AD**	
Age interval	Male	Female	Male	Female
$[50, 55)$	1	0	1	0
$[55, 60)$	5	4	5	4
$[60, 65)$	3	6	3	6
$[65, 70)$	6	10	6	10
$[70, 75)$	6	8	6	8
$[75, 80)$	3	2	3	2
Total	24	30	24	30

The evaluation metrics for the AD classification task are: $Accuracy = \frac{TN+TP}{N}$, Precision $\pi = \frac{TP}{TP+FP}$, Recall $\rho = \frac{TP}{TP+FN}$ and F-1 score $F_1 = 2\frac{\pi * \rho}{\pi + \rho}$.

where N is the number of patients, TP, FP and FN are the number of true positives, false positives and false negatives, respectively.

[5]https://github.com/huggingface/transformers

5. Results

This section presents the results in the AD classification for both the leave-one-subject-out (LOSO) and test settings of the ADReSS challenge.

5.1. Results for LOSO setting

All the systems have been trained using leave-one-subject-out (LOSO) cross-validation strategy for measuring the generalization error. Therefore, models use 107 subjects as training data and are tested on the held-out subject.

Table 2 illustrates the individual results achieved by the three speech-based systems. These results show that the x-vectors and fluency based systems have similar performance regarding the accuracy, as well as Area Under Curve (AUC). The i-vector was the weakest model, being the only one with an accuracy under 70%, although it is still above the challenge baseline results [7].

Table 2: *AD classification results of the proposed speech-based systems (LOSO cross-validation).*

	class	Precision	Recall	F1 Score	Accuracy	AUC
i-vector	non-AD AD	0.7000 0.6724	0.6481 0.7222	0.6730 0.6964	0.6851	0.6798
x-vector	non-AD AD	0.6923 0.7906	0.8333 0.6296	0.7563 0.7010	0.7314	0.7568
fluency	non-AD AD	0.7272 0.7358	0.7407 0.7222	0.7339 0.7289	0.7314	0.7613

The results for both text-based systems are shown in Table 3. Concerning BERT-based model, the AUC in held-out set is 0.9078. For a selected threshold probability, we also obtain an accuracy of 0.8518, recall of 0.8333 and F1-score of 0.8490. For the linguistic model, the AUC in held-out set is 0.7510. For a selected threshold probability the accuracy is 0.7129, precision of 0.8965, recall of 0.4814 and F1-score of 0.6265.

Table 3: *AD classification results of the proposed text-based systems (LOSO cross-validation).*

	class	Precision	Recall	F1 Score	Accuracy	AUC
BERT model	non-AD AD	0.8392 0.8653	0.8703 0.8333	0.8545 0.8490	0.8518	0.9077
Linguistic model	non-AD AD	0.6455 0.8965	0.9444 0.4818	0.7669 0.6265	0.7129	0.7510

Moreover, three different score fusion strategies were carried out. In the first one (referred as Fusion I), the scores of every system were normalized by z-norm, and merged by a weighted sum fusion rule. Weights are chosen from the accuracy of each individual AD detection system. In the second one (referred as Fusion II), the same normalization strategy was used to first, using also a weighted sum, merge the speech-based and text-based system scores separately, and then these new scores were again merged by a new weighted sum. Finally, in the last one (referred as Fusion III) instead of use text-based fused scores, only the BERT-based scores were fused with the speech-based fused scores. Figure 2 shows the ROC curves of the described fusion systems. The Fusion I model had an accuracy of 0.8611, a F1-score of 0.8543, and an AUC of 0.9355. The Fusion II model presented an accuracy of 0.8611, a F1-score of 0.8543 and an AUC of 0.9372. Lastly, the Fusion III model presented an accuracy of 0.8796, a F1-score of 0.8807

and an AUC of 0.9405. The results show that the fusion of both text-based and speech-based modalities improves the detection of AD.

For further insight, Figures 1 and 2 compares the ROC curves of all individual systems and their fusion, respectively.

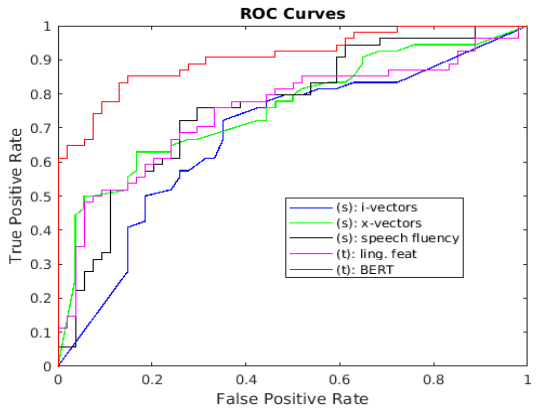


Figure 1: *ROC curves of all individual systems.*

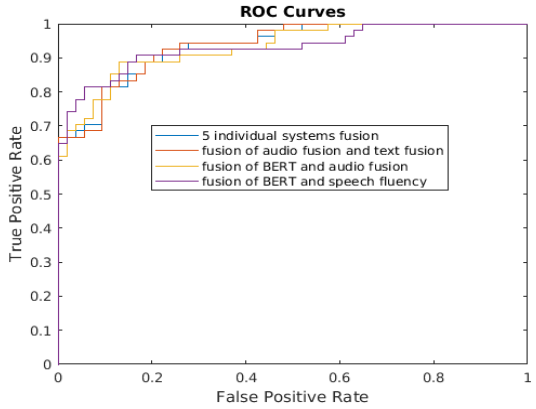


Figure 2: *ROC curves for the different fusion strategies.*

5.2. AD classification results for test setting.

Based on the LOSO cross-validation results the three best individual systems and all the fusion strategies were evaluated on the testing dataset. Table 4 shows the performance achieved by them on the testing dataset, which consist of recordings of 48 subjects.

The results achieved by the BERT model were very similar in validation and testing, which is a good sign of the lack of overfitting. However, the submitted speech-based systems had a significant decrease in performance. This fact shows that acoustic systems, require more data in order to improve the predictive ability of speech embeddings, and thus improve their generalization capacity avoiding overfitting issues. As a result, the performance of the fusion I and fusion II strategies also declines, but fusion III improves the accuracy of the BERT model.

6. Conclusions and Further work

Two lines of work have been developed on the ADReSS Challenge dataset: one based on speech processing and the other based on text processing.

Table 4: *Results on the testing dataset*

	class	Precision	Recall	F1 Score	Accuracy
Fusion I	non-AD AD	0.8182 0.7692	0.7500 0.8333	0.7826 0.8000	0.7917
Fusion II	non-AD AD	0.8000 0.8260	0.8333 0.7917	0.8163 0.8085	0.8125
Fusion III	non-AD AD	0.8636 0.8077	0.7917 0.8750	0.8261 0.8400	**0.8333**
BERT model	non-AD AD	0.7778 0.8571	0.8750 0.7500	0.8235 0.8000	0.8125
fluency	non-AD AD	0.6250 0.6250	0.6250 0.6250	0.6250 0.6250	0.6250
x-vector	non-AD AD	0.5417 0.5417	0.5417 0.5417	0.5417 0.5417	0.5417

It is important to highlight the following points when assessing both solutions:

- **Performance:** x-vector and fluency speech features have shown a competitive performance in the LOSO cross-validation. However, their performance decreases on the test setting. This would be a result of a low generalization level achieved at the training stage. On the other hand, the text-based BERT model has obtained superior results both in the LOSO and testing settings. The Fusion III strategy improves the accuracy of the BERT model and also obtains more balanced performance measures. Then, multimodal systems demonstrate to be a better strategy for the detection of AD than individual systems.
- **Complexity:** The x-vector and BERT systems, as deep-learning-based models, need a large amount of training data to be able to generalize well.
- **Explainability:** deep learning models are black-box models, being very difficult to interpret from a human perspective. On the contrary, the linguistic and fluency features are explainable and one could determine the weight of them in the classification.

Several future research lines have been identified for further work. Firstly, investigation on improved classification based on deep neural networks and novel acoustical feature extraction algorithms. Secondly, addition of new linguistic features and non-verbal information (breaks, silence duration, word mistakes, etc.) in text-based systems. Thirdly, analysis of strategies for increasing the generalization capacity of the proposed speech-based systems; for example: to find new rhythmic parameters more discriminatory between patients with and without AD or to adapt the deep learning models to the characteristics of elderly speech. It is also important to conduct experiments in other experimental settings, for example using other questions of the mini-mental state examination test, to validate the results obtained and, above all, to increase the size of the data settings.

7. Acknowledgements

This work has received financial support from the Spanish "Ministerio de Economia y Competitividad" through the project Speech&Sign RTI2018-101372-B-100, and also from Xunta de Galicia (AtlanTTic and ED431B 2018/60 grants) and European Regional Development FundERDF.

8. References

[1] A. Association, "2019 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 15, no. 3, pp. 321–387, 2019.

[2] P. J. Nestor, P. Scheltens, and J. R. Hodges, "Advances in the early detection of alzheimer's disease," *Nature medicine*, vol. 10, no. 7, pp. S34–S41, 2004.

[3] S. Ahmed, A.-M. F. Haigh, C. A. de Jager, and P. Garrard, "Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease," *Brain*, vol. 136, no. 12, pp. 3727–3737, 2013.

[4] J. J. G. Meilán, F. Martínez-Sánchez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana, "Speech in alzheimer's disease: Can temporal and acoustic parameters discriminate dementia?" *Dementia and Geriatric Cognitive Disorders*, vol. 37, no. 5-6, pp. 327–334, 2014.

[5] K. Lopez-de Ipiña, J. B. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, C. M. Travieso, M. Ecay-Torres, P. Martinez-Lage, and H. Eguiraun, "On automatic diagnosis of alzheimer's disease based on spontaneous speech analysis and emotional temperature," *Cognitive Computation*, vol. 7, no. 1, pp. 44–55, 2015.

[6] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[7] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," in *Proceedings of INTERSPEECH 2020*, Shanghai, China, 2020. [Online]. Available: https://arxiv.org/abs/2004.06833

[8] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 2162–2166. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-2516

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[10] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[12] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2014, pp. 2494–2498.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," 2011.

[14] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech and Language*, vol. 60, 2020.

[15] M. Ajili, S. Rossato, D. Zhang, and J.-F. Bonastre, "Impact of rhythm on forensic voice comparison reliability," in *Odyssey 2018: The Speaker and Language Recognition Workshop*. ISCA, 2018, pp. 1–8.

[16] A. Larcher, K. A. Lee, and S. Meignier, "An extensible speaker identification sidekit in python," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5095–5099.

[17] B. Macwhinney, "The childes project: tools for analyzing talk," *Child Language Teaching and Therapy*, vol. 8, 01 2000.

[18] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations (ICLR 2019)*. OpenReview.net, 2019.

[20] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's disease : JAD*, vol. 49, no. 2, pp. 407–422, 2016.

List of Authors