Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome

Mario Corrales-Astorgano[a,*], David Escudero-Mancebo[a], César González-Ferreras[a]

[a]*Departamento de Informática, Universidad de Valladolid, Valladolid, Spain*

Abstract

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 is 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 is 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 has been 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.

Keywords: Speech characterization, Prosody, Down syndrome, Intellectual disabilities, Automatic classification, Perceptual test

1. Introduction

Individuals with Down syndrome (DS) have problems in their language development that make their social relationships and their developmental ability more problematic (Cleland et al., 2010; Martin et al., 2009; Chapman, 1997). Many DS individuals have some physiological peculiarities that affect their voice production, such as a smaller vocal tract with respect to the tongue size or soft palatal shape, among others Guimaraes et al. (2008). Muscular hypotonia also affects their capabilities for performing a correct articulation, degrading the quality of the spectral characteristics of sounds (Markaki and Stylianou, 2011). In addition, hearing loss during childhood (Shott et al., 2001) and fluency deficits (Devenny and Silverman, 1990) influence the frequency, energy and temporal domains of the voice signal.

Although problems derived from physiological peculiarities are permanent (only surgery (Leshin, 2000) or prostheses (Bhagyalakshmi et al., 2007) could improve them), intonation and fluency deficits can be improved by speech therapy and training. There are tools available for this goal (González-Ferreras et al., 2017) based on perception and production activities to be performed with the assistance of therapists who help patients to properly manage their breathing and intonation patterns. Although there is general consensus about the importance of improving prosody by training (see (Kent and Vorperian, 2013) for a complete state of art revision), there are very few works that provide empirical evidence of the importance of the prosody related features (those belonging to fundamental frequency, energy and duration domains) with respect to other acoustic features belonging to the spectral domain.

The use of the video game described by González-Ferreras et al. (2017) has allowed the formation of a speech corpus, which has been used in this work to analyze and characterize the speech of people with Down syndrome. This corpus, described in section 3.1, contains recordings of people with Down syndrome and typically developing people. Both groups recorded the same sentences, so statistical and perceptual tests have been used to compare the acoustic features of the two groups of speakers, so that the most relevant differences could be identified.

This work aims to find the best acoustic features to characterize the speech of people with Down syndrome. To do this, features of frequency, energy, temporal and spectral domains have been extracted from the recordings

[*]Corresponding author

Email addresses: mcorrales@infor.uva.es (Mario Corrales-Astorgano), descuder@infor.uva.es (David Escudero-Mancebo), cesargf@infor.uva.es (César González-Ferreras)

of the gathered corpus. In addition, the relative weight of each domain in the characterization of people with Down syndrome has been included in this paper, especially the comparison between the spectral and the other domains.

The methodology described above was developed to answer two main research questions (RQ):

RQ1 : Which are the most discriminative acoustic features between the recordings of speakers with Down syndrome and typically developing speakers?

Issue 1.1 : Are there statistical differences between these features?

Issue 1.2 : Are these differences in accordance with what is expected or described in the state of the art?

RQ2 : What is the relative weight of the spectral features in comparison with the rest of the domains?

Issue 2.1 What is the relative weight of the different features when identifying atypical speech using automatic classifiers?

Issue 2.2 What is the relative weight of the different domains when identifying atypical speech in a perceptual test?

The structure of the article is as follows. Section 2 reviews related works from the state of the art and presents the innovation of our proposal. Section 3 describes the experimental procedure, including the corpus description, the features extraction process, the automatic classification experiment and the perceptual test. Section 4 shows the statistical test results of the different domain features, the automatic classification results and the perceptual test results. Finally, section 5 describes the discussion and section 6 the conclusions.

2. Background and related work

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 adults, Lee et al. (2009), Rochet-Capellan and Dohen (2015), Albertini et al. (2010) and Corrales-Astorgano et al. (2016) found significantly higher F0 values in adults with Down syndrome as compared to adults without intellectual disabilities. In addition, Lee et al. (2009) and Seifpanahi et al. (2011) found lower jitter (frequency perturbations) in adult speakers with Down syndrome. As for energy, Albertini et al. (2010) found significantly lower energy values in adults with Down syndrome. Moreover, Saz et al. (2009) concluded that adults with Down syndrome had poor control over energy in stressed versus unstressed vowels. Albertini et al. (2010) found lower shimmer (amplitude perturbations) in male adults with Down syndrome than in adults without intellectual disabilities. Finally, temporal domain results depend on the unit of analysis employed. Saz et al. (2009) found that people with cognitive disorders presented an excessive variability in vowel duration, while Rochet-Capellan and Dohen (2015) and Bunton and Leddy (2011) reported longer durations of vowels in adults with Down syndrome. Albertini et al. (2010) discovered a lower duration of words in male adults with Down syndrome. 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 ((Van Borsel and Vandermeulen, 2008; Devenny and Silverman, 1990; Eggers and Van Eerdenbrugh, 2017)). These disfluencies can affect the speech rhythm of people with Down syndrome.

On the other hand, Zampini et al. (2016) indicated that children with Down syndrome had lower F0 than children without intellectual disabilities. Moura et al. (2008) found higher jitter in children with Down syndrome than children without intellectual disabilities. In terms of energy, Moura et al. (2008) indicated higher shimmer in children with Down syndrome than in children without intellectual disabilities.

The unit of analysis and the phonation tasks used by the researchers are different. Rochet-Capellan and Dohen (2015) used Vowel-Consonant-Vowel bysyllabes, Saz et al. (2009) and Albertini et al. (2010) recorded words, Rodger (2009) and Zampini et al. (2016) built these corpora using semi-spontaneous speech and Corrales-Astorgano et al. (2016) analyzed sentences. Lee et al. (2009) combined words, reading and natural speech. The majority of the studies are focused on the English language (Kent and Vorperian, 2013), but there are others focused on Italian (Zampini et al., 2016; Albertini et al., 2010), Spanish (Corrales-Astorgano et al., 2016; Saz et al., 2009), French (Rochet-Capellan and Dohen, 2015) or Farsi (Seifpanahi et al., 2011).

The use of spectral features to assess pathological voice has frequently been applied in the literature. Dibazar et al. (2006) used MFCCs and pitch frequency with a hidden Markov model (HMM) classifier for the assessment of normal versus pathological voice using one vowel as the unit of analysis. Markaki and Stylianou (2011) suggested the use of modulation spectra for the detection and classification of voice pathologies. Markaki and Stylianou (2010) created a method for the objective assessment of hoarse voice quality, based on modulation spectra, using a corpus of sustained vowels. The voice quality was evaluated using the long term average spectrum (LTAS) and alpha ratio by Leino (2009). Although these works do not refer to people with Down syndrome, they do refer to some aspects that appear in this kind of speakers and we refer to them in the discussion section.

Formant frequency and amplitude have also been studied in people with Down syndrome. A larger vowel space in people with Down syndrome was found by Rochet-Capellan and Dohen (2015), while other studies denoted a reduction of the vowel space in children (Moura et al., 2008) and adults (Bunton and Leddy, 2011). Moreover, the voice of people with Down syndrome showed significantly reduced formant amplitude intensity levels (Pentz Jr, 1987).

In order to compare our study with the state of the art, a summary of other similar studies is shown in Table 1. A description of the corpus employed by these studies is shown in Table 2. To the best of our knowledge, our study is one of the first to analyze some features from the frequency, energy, temporal and spectral domains together. These features were extracted from the same recordings, which can help in the study of the relative importance of each domain in the characterization of the speech of people with Down syndrome. The use of a standard feature set (extended Geneva Minimalistic Acoustic Parameter Set, eGeMAPS; detailed in section 3.2 and Appendix A) can reduce the extraction methodology dependence, which can make it easier to compare the results of different studies.

Perceptual studies show mixed results. Moura et al. (2008) 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. Moran and Gilbert (1982) judged the voice quality of adults with Down syndrome as hoarse. In addition, Rodger (2009) noted discrepancies between perceptual judgments of pitch level and acoustic measures of F0. In our study, we did not want to compare each acoustic measure with a perceptual judgment of the same feature. Our aim is the assessment of the domain relevance in the identification of a recording as being from a person with Down syndrome, using automatic classifiers and perceptual tests.

3. Experimental procedure

Figure 1 shows the experimental methodology that we have followed. Firstly, the speech corpus recorded by people with Down syndrome and by typically developing people was gathered. Secondly, acoustic features were extracted from all the recordings of each corpus and a statistical test to analyze the differences between groups was carried out. Finally, the automatic classification experiment was carried out, in which the features with significant differences were used.

3.1. Corpus collection

We developed a computer video game to improve the prosodic and communication skills of people with Down syndrome (González-Ferreras et al., 2017). This video game is a graphic adventure game where users have to use the computer mouse to interact with the elements on the screen, listen to audio instructions and sentences from the characters of the game, and record utterances using a microphone in different contexts. The video game was designed using an iterative methodology in collaboration with a school of special education located in Valladolid (Spain). The feedback provided by teachers of special education was complemented by research into the difficulties of this population to use information and communication technologies. They have some difficulties, such as attention deficit(Martínez et al., 2011), lack of motivation(Wuang et al., 2011), or problems with the short term memory (Chapman and Hesketh, 2001) that had to be taken into account when developing the video game. The game was developed for the Spanish language.

Inside the narrative of the game, some learning activities were included to practice communication skills. There are three different types of activities: comprehension, production and visual. Firstly, the comprehension activities are focused on lexical-semantic comprehension and on improving prosodic perception in specific contexts. Secondly, production activities are focused on oral production, so the players are encouraged by the game to train their speech, keeping in mind such prosodic aspects as intonation, expression of emotions or syllabic emphasis. At the beginning of these activities, the video game introduces the context where the sentence has to be said. Then, the game plays the sentence and the player must utter the sentence while it is shown on the screen. The production activities include affirmative, exclamatory and interrogative sentences. Finally, visual activities include other activities designed to add variety to the game and to reduce the feeling of monotony while playing.

The video game collected examples of sentences with different modalities (i.e. declarative, interrogative and exclamatory). Usually, the intonation patterns vary depending on the modality. Neutral declarative sentences usually end with a decline to a low tone, while total interrogatives end with an upgrade to a high pitch. On the other hand, partial interrogative sentences, which are characterized by an interrogative element at the beginning of the sentence, start with a high tone associated with that interrogative element and usually end with a fall. Finally, exclamatory sentences are usually a marked variation of the corresponding declarative, so the variation lies basically in such aspects as the intensity, volume and tonal range used by the speaker.

Moreover, the combination of different sentences allows the inclusion of inflections that indicate a particular segmentation in oral production. Depending on the context and speed of elocution, these inflections may correspond to a pause, which implies a silence and, normally, the end of the sentence, or a semi-pause, which implies an intonation change in the same sentence. For instance, one of the examples collected in the corpus includes the three modalities and forces the speaker to make a pause between sentences: *¡Hola! ¿Tienen lupas? Quería comprar una. (Hello! Do you have magnifiers? I wanted to buy one*). In other cases, the tonal inflection corresponds to a semi-pause involving no change of modality or silence: *¡Hasta luego, tío Pau! (See you later, uncle Pau!*). Thus, the combination of these types of inflection allows the col-

Author	**Group**	**Frequency**	**Duration**	**Loudness**
Rodger 2009	Adults and Children	No differences		
Zampini 2016	Children	Good control for linguistics low for pragmatics. Lower F0.		
Saz 2009	Adults and Children	Good control in pronounced vowels	Longer pronounced vowels Dispersed mispronounced vowels	Low control of intensity in unstressed vowels
Albertini 2010	Adults	Higher F0	Lower duration (only for men)	Lower energy.Shimmer lower (only men)
Rochet-Capellan 2015	Adults	Higher F0	Longer vowels	
Lee 2009	Adults	Smaller pitch range. Higher F0. Lower jitter.		
Corrales 2016	Adults	Higher F0 excursions	More pauses to complete turns	Different range

Table 1: Results of different studies in the state of the art

Author	**Group**	**Down syndrome**	**Control**	**Type**	**Size**	**Language**
Rodger 2009	Adults and Children	22	52	Semi spontaneous	5 picture descriptions per speaker	English
Zampini 2016	Children	9	12	Semi spontaneous	20 minutes per speaker	Italian
Saz 2009	Adults and Children	3	168	Words	9576 words (6 hours) Control 684 words (38 minutes) Down syndrome	Spanish
Albertini 2010	Adults	30	60	Words	NA	Italian
Rochet-Capellan 2015	Adults	8	8	Vowel-consonant-vowel	144 per speaker	French
Lee 2009	Adults	9	9	Vowel. Reading. Natural speech	3 vowels per speaker 1 reading per speaker 1 minute per speaker	English
Corrales 2016	Adults	18	20	Sentences	479 utterances	Spanish

Table 2: Description of the corpus used in the state of the art

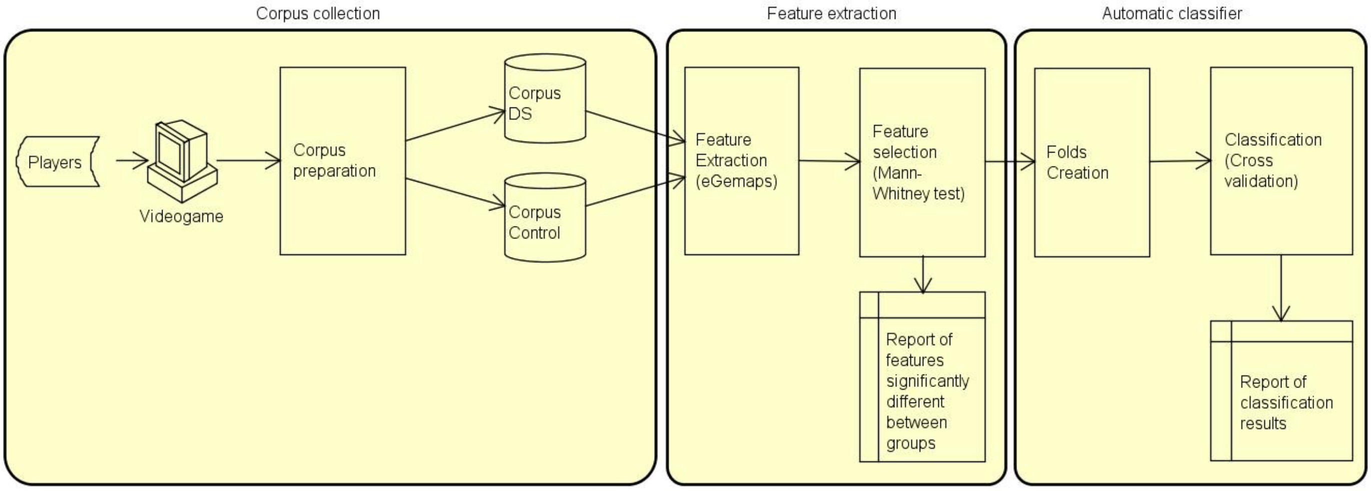
Corpus collection
Feature extraction
Automatic classifier
Players
Videogame
Corpus preparation
Corpus DS
Corpus Control
Feature Extraction (eGemaps)
Feature selection (Mann-Whitney test)
Report of features significantly different between groups
Folds Creation
Classification (Cross validation)
Report of classification results

Figure 1: Scheme of the experimental procedure which includes corpus collection, feature extraction and automatic classification

lection of examples with different segmentation. The sentences recorded can be seen in Table 3.

The recording sessions were carried out in the same facilities of the centers where the players attended their regular classes to assure the comfort of the players. In addition, a staff member of the centers was always with the players. The players were selected by the staff members because the distinct cognitive abilities of each student limited their possibilities as potential players, as some of them were not able to follow the structured process of the game in a reliable way. Eighteen speakers with Down syndrome participated, 11 males (chronological ages: 16, 16, 18, 20, 21, 21, 23, 24, 25, 26 and 30) and 7 females (chronological ages: 16, 17, 18, 19, 21, 22, 25). All of them were native speakers of Spanish, aged 16 to 30. They were students of two special education schools located in Valladolid and Barcelona(Spain) and have a moderate or mild intellectual disability. Besides, to reduce the ambient noise in the recording process, the players used a headset with a microphone incorporated (Plantronics USB headset). In addition, players recorded a different number of sentences, depending on their performance in the video game and the number of game sessions they did. It should be noted that for the production activities, not all speakers with Down syndrome reproduced the target sentence exactly. Some of them had hearing problems, while others had reading difficulties or cluttering derived from their intellectual disability.

To obtain a control sample of the recordings, twenty two adult speakers without any intellectual disability, 13 males and 9 females, were recorded. Therefore, two groups representing different populations were thus obtained: typically developing adults (TD) and people with Down syndrome (DS). Table 4 shows the number of users of each group of speakers, the number of recordings made by them and the total length in seconds of the recordings.

3.2. Feature extraction

Acoustic low-level descriptors (LLD) and temporal features were automatically extracted from each recording using the openSmile toolkit (Eyben et al., 2013). Two minimalistic feature sets were used. On the one hand, these sets provided enough features to characterize the audio recordings. On the other hand, the problem of having too many parameters relative to the number of observations. This problem can produce overfitting in the training phase, because the classifier adapts to the concrete set of inputs. This adaptation can produce good classification results for this particular set, but negatively affects the generalization capacity of the classifier. The Geneva Minimalistic Standard Parameter Set (GeMAPS) and the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), described by Eyben et al. (2016), were selected. The features extracted from each recording are sorted into four groups:

- Frequency related features: fundamental frequency and jitter.
- Energy related features: loudness, shimmer and Harmonics-to-Noise Ratio.
- Spectral features: alpha ratio, Hammarberg index, spectral slope, formant 1, 2, 3 relative energy, harmonic difference H1-H2, harmonic difference H1-A3, formant 1, 2, 3 frequency and formant 1, 2, 3 bandwidth.
- Temporal features: the rate of loudness peaks per second, mean length and standard deviation of continuous voiced and unvoiced segments and the rate of voiced segments per second, approximating the pseudo syllable rate.

In total, there are 25 LLD. The arithmetic mean and the coefficient of variation are calculated on these 25 LLD. Some functionals are applied to fundamental frequency and loudness: 20-th, 50-th, and 80-th percentile, the range of 20-th to 80-th percentile, and the mean and standard deviation of the slope of rising/falling signal parts. All these functionals are computed by the openSmile toolkit. In addition, the process used by the openSmile toolkit to extract the eGeMAPS features did not differentiate between silences and unvoiced regions, which can produce errors in the functions applied to each feature. Therefore, the Praat software (Boersma, 2006) was used to extract all silences from each recording and these silences were excluded from the analysis process.

Furthermore, 4 additional temporal features were added: the silence and sounding percentages, silences per second and the mean silences. These new features were added to improve the information about the temporal characterization of the recordings. In this case, the initial and final silence of each recording were excluded from the analysis process because their lengths were different due to the recording process. To sum up, the acoustic feature set contains 88 features from the eGeMAPS feature set and 4 new features introduced from the research team (92 features).

A statistical test was used to detect the significant differences between the features extracted from the recording of each group. The Mann-Whitney non-parametric test was used. Only the features with a p-value lower than 0.01 were selected for analysis and classification.

3.3. Automatic classification

In order to make an automatic classification of the recordings, the Weka machine learning toolkit (Hall et al., 2009) was used. This toolkit permits to a collection of machine learning algorithms to be accessed for data mining tasks. Three different classifiers were used to compare their performance: the C4.5 decision tree (DT), the multilayer perceptron (MLP) and the support vector machine (SVM).

Sentence in Spanish	**Sentence in English**
¡Hasta luego, tío Pau!	See you later, uncle Pau!
¡Muchas gracias, Juan!	Thank you very much, Juan!
¡Hola! ¿Tienen lupas? Quería comprar una.	Hello, do you have magnifiers? I wanted to buy one.
Sí, la necesito. ¿Cuánto vale?	Yes, I need it. How much is it?
¡Hola tío Pau! Ya vuelvo a casa.	Hello uncle Pau! I'll be back home.
Sí, esa es. ¡Hasta luego!	Yes, it is. Bye!
¡Hola, tío Pau! ¿Sabes dónde vive la señora Luna?	Hello uncle Pau! Do you know where Mrs Luna lives?
¡Nos vemos luego, tío Pau!	See you later, uncle Pau!
Has sido muy amable, Juan. Muchas gracias!	You have been very kind, Juan. Thank you very much!
¡Hola! ¿Tienen lupas? Me gustaría comprar una.	Hello, do you have magnifiers? I would like to buy one.
Sí, necesito una sea como sea. ¿Cuánto vale?	Yes, I really need one. How much is it?
Sí, lo es. Vivo allí desde pequeño. ¡Hasta luego!	Yes, it is. I have lived there since I was a child. Bye!
¡Hola, tío Pau! Tengo que encontrar a la señora Luna ¿Sabes dónde vive?	Hello uncle Pau! I have to find Mrs Luna. Do you know where she lives?

Table 3: Sentences included in the corpus

User type	#Users	#Recordings	Length(seconds)
Control (TD)	22	250	650
Down syndrome (DS)	18	349	1442

Table 4: Number of users and recordings of each group of the corpus

In addition, the 10-fold cross validation technique was used to create the training and testing datasets. To avoid classifier adaptation, all folds were created by recordings of different speakers. Therefore, the recordings of each speaker were joined in the same fold and each fold was balanced in terms of the number of recordings.

To analyze the performance of the classification, we used the classification rate. The unweighted average recall (UAR) (Schuller et al., 2016) 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.

3.4. Perception test

In order to evaluate the impact of prosody in the perception of the listeners, we used prosody transfer techniques. These techniques have previously been used in other studies of the state of the art. For instance, Luo et al. (2017) investigated the role of different prosodic features in the naturalness of English L2 speech. The prosodic modification method was applied to native and L2 learners' speech. Later, they used a perceptual test to evaluate the impact of prosody modification. A similar methodology was used by Escudero et al. (2017), where the characteristic prosodic patterns of the style of different groups of speakers was investigated. After the prosodic modification of the utterances, the characteristic prosodic patterns were validated using a perceptual test. The procedure described in Escudero et al. (2017) for transferring prosody is used in the experiments reported in this paper.

Figure 2 shows the experimental procedure used to perform the perception test. The sentence *¡Hola tío Pau! ¿Sabes donde vive la señora Luna?* (*Hello uncle Pau! Do you know where Mrs Luna lives?*) recorded by all the speakers was selected. This sentence was selected because of its prosodic richness (combining an affirmative and an interrogative sentence), because it was used in another of our studies (González-Ferreras et al., 2017) and because it was the most recorded sentence. To obtain a phonetic segmentation of the recordings, the BAS web services (Schiel, 1999; Kisler et al., 2017) were used. This tool returns the time intervals of each phoneme using the audio file and the transcription as inputs. Manual revision of the segmentation was necessary to correct transcription errors. The sentence was recorded by 22 TD speakers and by 16 speakers with DS. However, each speaker did not have the same number of recordings. In total, there were 62 recordings.

Once the segmentation was corrected, a prosody transfer algorithm implemented in Praat (Boersma, 2006) was executed. This algorithm transfers, phoneme by phoneme, 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. The algorithm was executed combining the audios of each speaker with the audios of the rest of the speakers,

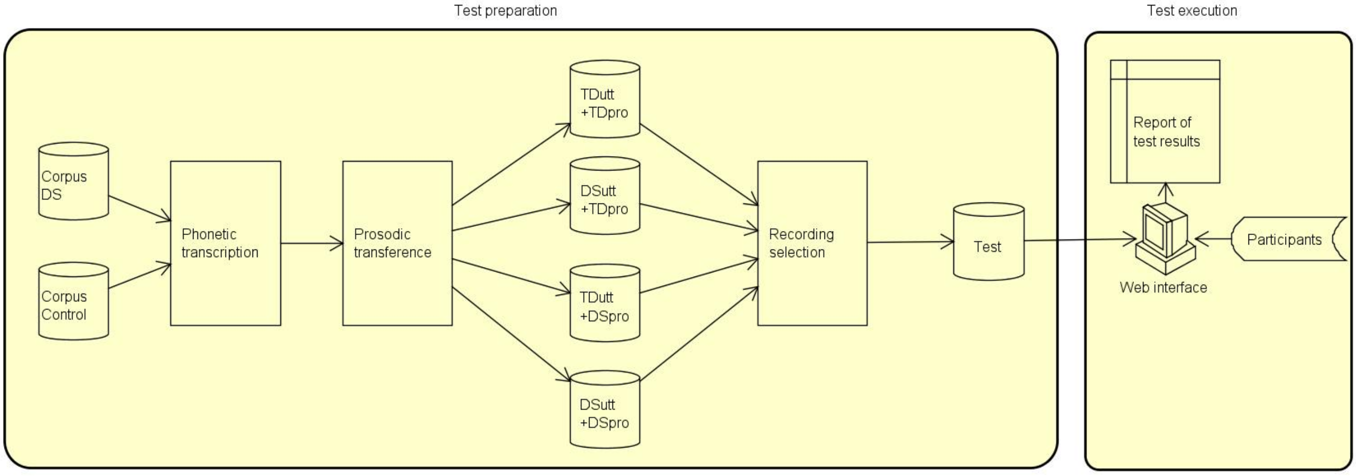

Figure 2: Experimental procedure followed to perform the perceptual test. The utterances used in the test were: TDutt+TDpro (utterance of a TD person with prosody transferred from an utterance of another TD person), DSutt+TDpro (utterance of a person with DS with prosody transferred from an utterance of a TD person), TDutt+DSpro (utterance of a TD person with prosody transferred from an utterance of a person with DS) and DSutt+DSpro (utterance of a person with DS with prosody transferred from an utterance of another person with DS).

so, in total, 3525 audio files were generated (not all the speakers had the same number of recordings). As a result, there are four types of audio files, as shown in Figure 2. Five audio files of each type were selected randomly for the perception test, so the test included twenty audio files, balanced in terms of gender.

The perception test was performed using a web application. First, personal information of the evaluator was collected. Then, the twenty audio files selected in the previous phase were shown randomly. The evaluators have to answer the following question for each utterance: *keeping in mind the way of speaking, do you think that the person who is speaking has intellectual disabilities? Ignore the audio distortion produced by the non natural voice synthesis.* The possible answers to the question were in a 5-point Likert scale: 1 means "no way" and 5 means "very sure". Thirty evaluators judged each utterance using this scale. People without any specific background on speech therapies were selected for this test, as we were interested in the perception of normal people concerning the importance of prosody in the identification of speech from people with intellectual disability.

4. Results

4.1. Characterization results

Table 5 shows the features with statistically significant differences (Mann-Whitney test with p-value ¡ 0.01) related to frequency, energy and temporal domains, sorted by mean differences. In the case of frequency, 9 of 12 features present significant differences. The first rows (from F0_stddevRisingSlope to jitter_mean) refer to the temporal evolution of the F0 contour. In all cases, figures present a higher value for speakers with Down syndrome, both when the *stddev* value is analyzed or the *Risingslope* and *jitter* (*jitter* value is lower because it focuses on the periods, which are the inverse of the F0 values). The last rows refer to mean values, coefficient of variation, ranges and percentiles of the F0 contour (from F0_pctlrange to F0_percetile80). Speakers with Down syndrome exhibit higher values than the speakers of the control group in all the cases, with a lower coefficient of variation in the Down syndrome group. These results seem to indicate that the participants with Down syndrome use higher F0 values with more temporal changes in the F0 contours.

There are 9 of 14 energy features that present statistically significant differences (Mann-Whitney test with p-value ¡ 0.01), as shown in Table 5. The first four rows (from loudness_percentil20 to loudness_pctlrange) refer to mean, range and percentile values. Values are higher for speakers with Down syndrome in all the cases. The last columns refer to the temporal variation of the energy values. In this case, Down syndrome speakers exhibit lower values. These results seem to indicate that participants with Down syndrome speak louder with less variation in the energy.

With respect to the temporal features displayed in Table 5, 9 of 10 features presented statistically significant differences (Mann-Whitney test with p-value ¡ 0.01). Speakers with Down syndrome use more pauses and they are longer (higher silencePercentage, silencePerSecond and silenceMean). The length of the voiced segment is longer, indicating that participants with Down syndrome speak more slowly.

As for spectral features (Table 6), 34 of 56 features showed statistically significant differences (Mann-Whitney test with p-value ¡ 0.01). Results show that the LTAS could be a useful instrument to detect differences, as clear differences appear when the features related with slope, Hammarberg and alpha index are taken into account. For-

Variable	Control	Control (CI 95%)	Down syndrome	Down syndrome (CI 95%)
F0 domain				
F0_stddevRisingSlope	166.17±231.44	(137.35,195.01)	220.85±273.67	(192.08,249.62)
jitter_stddevNorm	1.15±0.39	(1.11,1.21)	1.46±0.47	(1.42,1.52)
jitter_mean	0.04±0.02	(0.045,0.050)	0.03±0.01	(0.035,0.039)
F0_pctlrange	4.63±1.9	(4.4,4.88)	3.91±2.88	(3.61,4.22)
F0_percentile20	26.89±4.49	(26.33,27.45)	30.32±4.63	(29.84,30.81)
F0_percentile50	29.18±4.22	(28.66,29.71)	32.33±4.28	(31.89,32.79)
F0_mean	29.3±4.11	(28.79,29.82)	32.38±4.14	(31.95,32.82)
F0_stddevNorm	0.13±0.07	(0.129,0.147)	0.12±0.07	(0.116,0.132)
F0_percentile80	31.52±4.34	(30.99,32.07)	34.24±4.67	(33.75,34.73)
Energy domain				
loudness_percentile20	0.95±0.38	(0.91,1.01)	1.77±1.03	(1.66,1.88)
loudness_percentile50	1.93±0.73	(1.84,2.02)	3.29±2.22	(3.06,3.53)
loudness_mean	2.09±0.78	(1.99,2.19)	3.37±1.99	(3.17,3.58)
loudness_percentile80	3.15±1.24	(3,3.31)	4.9±2.94	(4.6,5.22)
loudness_pctlrange	2.19±0.96	(2.08,2.32)	3.13±2.06	(2.92,3.35)
loudness_stddevRisingSlope	15.3±7.18	(14.41,16.2)	19.63±14.24	(18.14,21.13)
loudness_stddevNorm	0.57±0.07	(0.57,0.58)	0.49±0.07	(0.48,0.5)
shimmer_mean	1.55±0.38	(1.51,1.61)	1.36±0.37	(1.32,1.4)
shimmer_stddevNorm	0.86±0.14	(0.84,0.88)	0.78±0.16	(0.77,0.8)
Temporal domain				
silencePercentage	0.1±0.11	(0.09,0.12)	0.22±0.19	(0.2,0.24)
silencesMean	0.16±0.2	(0.14,0.19)	0.31±0.3	(0.28,0.35)
StddevVoicedSegmentLengthSec	0.15±0.08	(0.14,0.16)	0.25±0.2	(0.23,0.27)
MeanVoicedSegmentLengthSec	0.26±0.15	(0.25,0.29)	0.44±0.39	(0.41,0.49)
silencesPerSecond	0.39±0.38	(0.35,0.44)	0.57±0.4	(0.53,0.62)
VoicedSegmentsPerSec	3.42±1.06	(3.29,3.55)	2.47±1.04	(2.37,2.59)
loudnessPeaksPerSec	5.76±1	(5.64,5.89)	4.39±0.94	(4.29,4.49)
MeanUnvoicedSegmentLength	0.05±0.02	(0.05,0.06)	0.06±0.03	(0.06,0.07)
soundingPercentage	0.89±0.11	(0.88,0.91)	0.77±0.19	(0.76,0.8)

Table 5: List of frequency, energy and temporal features with higher statistically significant differences (Mann-Whitney test with p-value $<$ 0.01), sorted by mean differences. The meaning of the features in column *Variable* can be seen in Appendix A. The units are reported in (Eyben et al., 2016).

Variable	Control	Control (CI 95%)	Down syndrome	Down syndrome (CI 95%)
LTAS related features				
slopeV0500_mean	0±0.03	(0,0.01)	0.05±0.03	(0.056,0.063)
slopeUV0500_mean	-0.06±0.04	(-0.07,-0.06)	0.05±0.03	(0.02,0.03)
slopeV0500_stddevNorm	-1.12±13.82	(-2.85,0.6)	0.69±2.64	(0.41,0.97)
alphaRatioUV_mean	-12.06±11.37	(-13.48,-10.65)	1.07±6.37	(0.41,1.75)
hammarbergIndexUV_mean	20.79±13.51	(19.11,22.48)	5.4±7.24	(4.64,6.16)
alphaRatioV_mean	-11.79±5.52	(-12.49,-11.11)	-8.46±5.55	(-9.05,-7.88)
hammarbergIndexV_mean	20.8±7.06	(19.93,21.69)	16.35±7.14	(15.61,17.11)
hammarbergIndexV_stddevNorm	0.48±0.67	(0.4,0.57)	0.57±1.01	(0.47,0.68)
slopeV5001500_mean	-0.02±0	(-0.03,-0.02)	-0.02±0	(-0.021,-0.020)
spectralFlux_mean	1.96±1.09	(1.83,2.1)	2.94±2.32	(2.7,3.19)
spectralFluxUV_mean	1.4±1.35	(1.23,1.57)	2.1±2.11	(1.88,2.32)
spectralFluxV_mean	2.11±1.12	(1.98,2.26)	3.13±2.53	(2.87,3.4)
spectralFlux_stddevNorm	0.72±0.19	(0.7,0.75)	0.67±0.12	(0.66,0.69)
MFCC related features				
mfcc3_stddevNorm	0.25±24.92	(-2.85,3.36)	-54.35±1039.94	(-163.68,54.98)
mfcc2V_mean	1.49±7.41	(0.58,2.42)	-2.45±6.88	(-3.17,-1.73)
mfcc4_stddevNorm	1.54±44.52	(-4.01,7.09)	-2±19.36	(-4.04,0.03)
mfcc2_stddevNorm	1.97±26.17	(-1.29,5.23)	-1.16±27.11	(-4.01,1.69)
mfcc2_mean	4.05±7.08	(3.18,4.94)	-2.32±6.45	(-3,-1.64)
mfcc4V_stddevNorm	-1.23±9.51	(-2.42,-0.05)	-0.45±4.73	(-0.96,0.04)
mfcc4_mean	-11.17±7.74	(-12.14,-10.21)	-17.34±9.91	(-18.39,-16.3)
mfcc3V_stddevNorm	-0.78±71.43	(-9.68,8.11)	-0.28±21.18	(-2.51,1.94)
mfcc4V_mean	-14.75±8.58	(-15.82,-13.68)	-18.3±10.83	(-19.44,-17.17)
mfcc1V_mean	26.42±7.31	(25.51,27.34)	20.93±9.61	(19.93,21.95)
mfcc1_mean	22.52±7.73	(21.56,23.49)	18.16±9.95	(17.11,19.21)
Formants related features				
F3amplitudeLogRelF0_stddevNorm	-1.18±0.25	(-1.22,-1.16)	-1.36±0.41	(-1.41,-1.32)
F2amplitudeLogRelF0_mean	-49.47±17.65	(-51.68,-47.28)	-42.63±20.55	(-44.79,-40.47)
F2amplitudeLogRelF0_stddevNorm	-1.35±0.26	(-1.39,-1.32)	-1.54±0.61	(-1.61,-1.48)
F1bandwidth_stddevNorm	0.2±0.08	(0.19,0.21)	0.23±0.09	(0.22,0.24)
F1frequency_stddevNorm	0.35±0.09	(0.34,0.37)	0.4±0.09	(0.39,0.41)
F3frequency_stddevNorm	0.09±0.02	(0.095,0.102)	0.1±0.02	(0.1,0.11)
F3frequency_mean	2665.98±145.97	(2647.81,2684.17)	2643.51±203.27	(2622.15,2664.89)
F3amplitudeLogRelF0_mean	-53.64±17.44	(-55.82,-51.47)	-45.02±19.5	(-47.08,-42.98)
Harmonic differences features				
logRelF0H1A3_stddevNorm	1.6±16.02	(-0.39,3.6)	0.18±7.44	(-0.6,0.97)
logRelF0H1A3_mean	18.91±6.26	(18.13,19.69)	15.86±7.09	(15.12,16.61)

Table 6: List of spectral features with higher statistically significant differences (Mann-Whitney test with p-value < 0.01), sorted by mean differences. The meaning of the features in column *Variable* can be seen in Appendix A. The units are reported in (Eyben et al., 2016).

mant 1 and Formant 3 (to a lower degree) also allow differences to be identified. As expected, MFCC values (the four analyzed) permit both groups to be separated. With respect to the variables related with the harmonic differences, only two variables appear in the list: logRelF0H1A3_stddevNorm and logRelF0H1A3_mean.

4.2. Classification results

Table 7 shows the classification results in the task of identifying the group of the speaker (TD or SD) of each utterance. The classifiers explained in section 3.3 and the selected features presented in the previous section were used. Only the features with significant differences between TD and DS groups are used. DT shows the lower classification results in all feature groups. MLP shows a better performance using frequency (UAR 0.64), temporal (UAR 0.78), frequency+energy+temporal (UAR 0.91) and all (UAR 0.95) feature groups. SVM works better with energy features (UAR 0.78). The results using spectral features are the same in MLP and SVM classifiers (UAR 0.87).

In addition, the best classification results are obtained using all features, independently of which classifier is used. Frequency features show the worst performance when they are used alone. Energy and temporal features have similar results, with only 9 features per group.

When frequency, energy and temporal features are used together, the performance is noticeably better than using each group separately. Finally, spectral features show a slightly worse performance than all and frequency+energy +temporal features.

4.3. Perception test results

Table 8 shows the results of the perception test and Figure 3 visually presents the differences between the groups. When the prosody of TD speakers was transferred to utterances of TD speakers, 84% of the answers identified the audios as TD speakers (answer 1 of row TDutt+TDpro). In this case, the doubts in the identification of the audio files as TD or DS represent only 2% of the answers (answer 3 of row TDutt+TDpro). On the other hand, when the prosody of DS speakers was transferred to utterances of DS speakers, 73% of the answers identified the audios as DS speakers (answers 4 and 5 of row DSutt+DSpro). In this case, the doubts in the identification of the audio files as TD or DS represent 18% of the answers (answer 3 of row DSutt+DSpro), and the identifications as TD are only 8% (answers 1 and 2 of row DSutt+DSpro).

The answers given about the audio files that combined utterances of one group with prosody of the other group present much more variability. However, prosody had more influence in the identification process than the original utterance. When the prosody of TD speakers was transferred to utterances of speakers with DS, 58% of the answers identified the audios as TD speakers (answers 1 and 2) versus only 20% of DS identifications (answers 4 and 5). On the other hand, 51% of the answers identified the audios as speakers with DS (answers 4 and 5) when the prosody of speakers with DS was transferred to an utterance of TD speakers, versus only 26% of TD identifications (4 and 5 answers). In both cases, the number of answers 3 is relevant (22% and 23% of answers 3, respectively).

Moreover, two statistical tests were used to compare the answers obtained. The results of the Kruskal-Wallis non-parametric test showed significant differences (with a p-value < 0.001) between the answers given to the four groups (TDutt+TDpro, DSutt+TDpro, TDutt+DSpro and DSutt+DSpro). Furthermore, the Mann-Whitney non-parametric test was used to compare each group with the others, in groups of two. All the comparisons showed significant differences (p-value < 0.001).

5. Discussion

5.1. Characterization of the speech of people with Down syndrome

Fundamental frequency is significantly higher in speakers with Down syndrome. The same results were found by Albertini et al. (2010), Rochet-Capellan and Dohen (2015) and Lee et al. (2009). In addition, the F0 range is lower in speakers with Down syndrome, which can be explained by a less melodious intonation. Continuing with frequency, jitter is significantly lower in the DS group, as found by Lee et al. (2009) and by Seifpanahi et al. (2011).

Concerning temporal features, on the one hand, the number of continuous voiced regions per second is lower in the speakers with Down syndrome, which means that the oral production of speakers with Down syndrome was slower than that of control speakers. Reading difficulties that some people with Down syndrome present can have influenced these results. On the other hand, Van Borsel and Vandermeulen (2008) found disfluencies in Down syndrome speaking, such as cluttering and stuttering. These disfluencies can produce the insertion of more silences and the presence of more temporal variety in the speech of people with Down syndrome, as found in this study.

In terms of energy, loudness features were found to be significantly higher in the speakers with Down syndrome and its range was higher. This result contradicts that reported by Albertini et al. (2010), which showed lower energy values in speakers with Down syndrome. Another study focused on vowels (Saz et al., 2009) found an increase in the energy of unstressed vowels in Down syndrome speakers. Energy is always a difficult variable in the analysis of prosody, as its values are very dependent on the recording conditions: the dynamic range of the microphone and the distance between the speaker and the microphone. On the other hand, some of the participants have slight hearing problems, which may be another possible explanation for the higher energy values.

Our corpus also permitted the detection of differences related with the spectral features. Table 6 highlights the

		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	72.5	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	86.5	0.87

Table 7: 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.

Type	**1**	**2**	**3**	**4**	**5**	**NR**	**Total**
TDutt+TDpro	124	15	3	1	4	3	150
DSutt+TDpro	42	42	31	18	11	6	150
TDutt+DSpro	17	21	34	43	31	4	150
DSutt+DSpro	1	11	26	49	56	7	150

Table 8: Number of responses of the perception tests for each type of audio file. A response of 1 means "no way" and 5 means "very sure" in the identification of the audio file as a speaker with Down syndrome. NR means no response. TDutt+TDpro means utterance of a TD person with prosody transferred from an utterance of another TD person; DSutt+TDpro means utterance of a person with DS with prosody transferred from an utterance of a TD person; TDutt+DSpro means utterance of a TD person with prosody transferred from an utterance of a person with DS; and DSutt+DSpro means utterance of a person with DS with prosody transferred from an utterance of another person with DS.

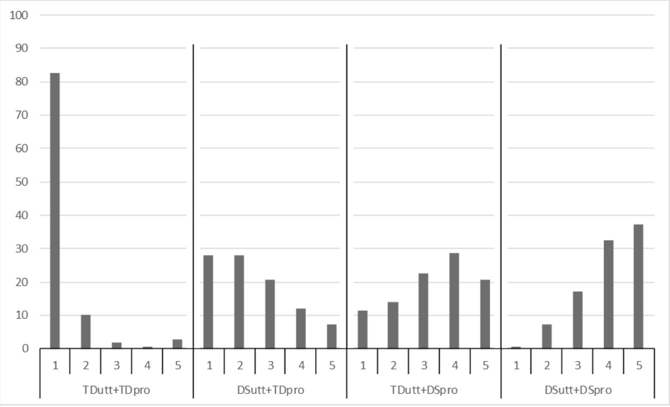

Figure 3: Results of the perception tests for each type of audio file. TDutt+TDpro means utterance of a TD person with prosody transferred from an utterance of another TD person; DSutt+TDpro means utterance of a person with DS with prosody transferred from an utterance of a TD person; TDutt+DSpro means utterance of a TD person with prosody transferred from an utterance of a person with DS; and DSutt+DSpro means utterance of a person with DS with prosody transferred from an utterance of another person with DS.

fact that LTAS has been proposed in Gauffin and Sundberg (1977) for the identification of breathy and hypokinetic voice. The relative amplitude of the first harmonic was also related with breathy voices by Hillenbrand and Houde (1996). The speech of people with DS is described as breathy by Wold DC (1979) and dysphonic by Moran (1986). MFCC features are commonly used in speaker recognition applications (Martinez et al., 2012), as they are representative of the vocal tract shape (Dusan and Deng, 1998). The relative importance of the MFCC features on the characterization of the speech of people with DS (as shown in Table 6) could thus be justified by the special anatomy of the tongue, palate, jaw, etc. of this type of speaker (Rodger, 2009). MFCC has also been used to identify nasality by Yuan and Liberman (2011) which is another aspect that has been related with the speech of people with DS in many works (Kent and Vorperian, 2013). The relative position of the formants has been associated with the degree of nasality in many works (House and Stevens, 1956; Huffman, 1989) which was also highlighted in our results table.

Finally, people with DS present hypotonia of muscles and difficulties in motor control, which affect the movement of the lips, tongue and jaw, with the consequent impact on spectral features already mentioned. The lack of muscular strength could also be another reason justifying the slower speech. As hypotonia could also affect the diaphragm, the energy values should have been lower. We hypothesize that the reason why higher values of energy were obtained could be due to the extra effort made by students to correctly complete the activities.

5.2. Relative impact of prosody

The experimental results obtained show that the features concerning the frequency, energy and temporal domains have the same or a greater impact than the spectral domain features to identify the speech of people with Down syndrome:

- There are a high number of features out of the spectral domain that present significant differences between speakers with Down syndrome and speakers without intellectual disabilities.

- Spectral features achieve high classification rates (up to 87%), but classification rates of frequency, energy and temporal features together are higher than spectral features (up to 91.83%).

- Utterances of control speakers with transferred frequency, energy and phoneme duration from speakers with Down syndrome are mostly perceived as anomalous voice. In the same way, utterances of speakers with Down syndrome with transferred frequency, energy and phoneme duration from control speakers are mostly perceived as typical speech.

To the best of our knowledge, there are few studies that assess, in an experimental way, the relative weight of prosody in the perception of speech of people with Down syndrome as a non typical voice. The differences between speakers with Down syndrome and control speakers in the spectral domain can be derived from physiological peculiarities in their phonological system. Some could be corrected by surgery, but others are impossible to be corrected. However, frequency, energy and temporal characteristics can be trained using speech therapy techniques focusing on breathing and repetition of activities. The results obtained in this paper show the potential benefits of prosody training.

The distance between the prosodic features of speakers with Down syndrome and those of control speakers can be used to devise a quality metric to be included in computer assisted pronunciation training applications. Our future work on the implementation of an automatic evaluation module of voice quality is expected to benefit from the results of this paper. This module is to be included in our speech training tools (González-Ferreras et al., 2017), so spectral features will be useful to identify a recording as a non typical speech, while prosody analysis will be necessary for the evaluation of the players' improvement over the different game sessions.

5.3. Limitations

The corpus size in speech analysis studies is very important to achieve representative results. The recording of a corpus of speech of people with Down syndrome is always challenging because of the special characteristics of these speakers (attention deficit and problems with short term memory, among others). Our video game has allowed the recording of a speech corpus whose size is bigger than other speech corpora used in other studies (see Table 2). Although the corpus size could be larger, the statistical tests carried out guarantee that the corpus has the necessary size to obtain significant results. In addition, new recordings are currently being obtained due to the use of the video game in a school of special education.

The heterogeneity of the population with Down syndrome can have an influence on the correct generalization of the results. However, the methodology presented in this paper can be applied to individuals with the aim of identifying the concrete features that they are using wrongly. Moreover, the relative impact of these features in the identification of their speech as pathological can be analyzed.

6. Conclusions

The speech characterization experiment presented in this article has allowed us to find significant differences between the speech of individuals with Down syndrome and those of the control group that affect the use of a set of acoustic variables related to frequency, energy, temporal and spectral domains. The use of these variables in an experiment of automatic identification allows very high classification rates (above 95%) to be obtained. If these variables are used independently, the classification rates decrease, the highest being those obtained using the spectral features. However, the importance of the rest of the variables becomes clear, because when only the variables related to frequency, energy and temporal domains are used, the classification rate can be higher than that obtained using the spectral features.

A perception experiment, based on prosody transfer, allowed us to verify the high relative importance of the prosodic variables of frequency, energy and temporal domains regarding the perception of atypical speech. An adequate control of these variables in utterances of speakers with Down syndrome allows us to change the perception of them, even though the voice quality is not modified. Besides, transferring the prosody from speakers with Down syndrome to speakers of the control group means the utterances will be perceived, to a large degree, as if they were from speakers with Down syndrome. This result encourages the use of methodologies for training prosody as a means for improving the overall quality of the oral production of Down syndrome speakers.

Acknowledgments

The work described in this paper was supported (1/2016-12/2017) by the Fundacion BBVA (project "Pradia: la aventura gráfica de la pragmática y la prosodia" - CF613399). The activities of Down syndrome speech analysis continue (1/2018-12/2020) in the project funded by the Ministerio de Economía, Industria y Competitividad (MINECO) and the European Regional Development Fund FEDER (project "Incorporación de un Módulo de Predicción Automática de la Calidad de la Comunicación Oral de Personas con Síndrome de Down en un Videojuego Educativo" - TIN2017-88858-C2-1-R). The authors would like to thank all the participants who took part in the recording of the corpus. We would also like to thank Lourdes Aguilar, Valle Flores, Yolanda Martín and Ferran Adell. Special thanks to the students of the Fundación Personas (http://www.fundacionpersonas.org) for their motivation during the training sessions.

References

Albertini G, Bonassi S, Dall'Armi V, Giachetti I, Giaquinto S, Mignano M. Spectral analysis of the voice in Down syndrome. Research in developmental disabilities 2010;31(5):995–1001.

Bhagyalakshmi G, Renukarya A, Rajangam S. Metric analysis of the hard palate in children with down syndrome-a comparative study. Down Syndrome Research and Practice 2007;12(1):55–9.

Boersma P. Praat: doing phonetics by computer. http://www praat org/ 2006;.

Bunton K, Leddy M. An evaluation of articulatory working space area in vowel production of adults with Down syndrome. Clinical linguistics & phonetics 2011;25(4):321–34.

Chapman R, Hesketh L. Language, cognition, and short-term memory in individuals with Down syndrome. Down Syndrome Research and Practice 2001;7(1):1–7.

Chapman RS. Language development in children and adolescents with Down syndrome. Mental Retardation and Developmental Disabilities Research Reviews 1997;3(4):307–12.

Cleland J, Wood S, Hardcastle W, Wishart J, Timmins C. Relationship between speech, oromotor, language and cognitive abilities in children with Down's syndrome. International journal of language & communication disorders 2010;45(1):83–95.

Corrales-Astorgano M, Escudero-Mancebo D, González-Ferreras C. Acoustic analysis of anomalous use of prosodic features in a corpus of people with intellectual disability. In: Advances in Speech and Language Technologies for Iberian Languages: Third International Conference IberSPEECH. Springer; 2016. p. 151–61.

Devenny D, Silverman W. Speech dysfluency and manual specialization in Down's syndrome. Journal of Intellectual Disability Research 1990;34(3):253–60.

Dibazar AA, Berger TW, Narayanan SS. Pathological voice assessment. In: Engineering in Medicine and Biology Society (EMBS). IEEE; 2006. p. 1669–73.

Dusan S, Deng L. Recovering vocal tract shapes from mfcc parameters. In: ICSLP. 1998. .

Eggers K, Van Eerdenbrugh S. Speech disfluencies in children with Down Syndrome. Journal of Communication Disorders 2017;.

Escudero D, González C, Gutiérrez Y, Rodero E. Identifying characteristic prosodic patterns through the analysis of the information of Sp_ToBI label sequences. Computer Speech & Language 2017;45:39–57.

Eyben F, Scherer KR, Schuller BW, Sundberg J, André E, Busso C, Devillers LY, Epps J, Laukka P, Narayanan SS, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing 2016;7(2):190–202.

Eyben F, Weninger F, Gross F, Schuller B. Recent developments in opensmile, the Munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM international conference on Multimedia. ACM; 2013. p. 835–8.

Gauffin J, Sundberg J. Clinical applications of acoustic voice analysis. Part II: acoustical analysis, results, and discussion. Speech Transmission Laboratory, Quarterly Progress and Status Report 1977;2:39–43.

González-Ferreras C, Escudero-Mancebo D, Corrales-Astorgano M, Aguilar-Cuevas L, Flores-Lucas V. Engaging adolescents with Down syndrome in an educational video game. International Journal of Human–Computer Interaction 2017;:1–20.

Guimaraes CV, Donnelly LF, Shott SR, Amin RS, Kalra M. Relative rather than absolute macroglossia in patients with Down syndrome: implications for treatment of obstructive sleep apnea. Pediatric radiology 2008;38(10):1062.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The weka data mining software: an update. ACM SIGKDD explorations newsletter 2009;11(1):10–8.

Hillenbrand J, Houde RA. Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. Journal of Speech, Language, and Hearing Research 1996;39(2):311–21.

House AS, Stevens KN. Analog studies of the nasalization of vowels. Journal of Speech and Hearing Disorders 1956;21(2):218–32.

Huffman MK. Implementation of nasal: timing and articulatory landmarks. Ph.D. thesis; University of California, Los Angeles; 1989.

Kent RD, Vorperian HK. Speech impairment in Down syndrome: a review. Journal of Speech, Language, and Hearing Research 2013;56(1):178–210.

Kisler T, Reichel U, Schiel F. Multilingual processing of speech via web services. Computer Speech & Language 2017;45:326–47.

Lee MT, Thorpe J, Verhoeven J. Intonation and phonation in young adults with Down syndrome. Journal of Voice 2009;23(1):82–7.

Leino T. Long-term average spectrum in screening of voice quality in speech: untrained male university students. Journal of Voice 2009;23(6):671–6.

Leshin L. Plastic surgery in children with down syndrome. Down syndrome: Health issues: News and information for parents and professionals 2000;.

Luo D, Luo R, Wang L. Prosody analysis of L2 English for naturalness evaluation through speech modification. In: Proc. Interspeech. 2017. p. 1775–8.

Markaki M, Stylianou Y. Modulation spectral features for objective voice quality assessment. In: Communications, Control and Signal Processing (ISCCSP), 2010 4th International Symposium on. IEEE; 2010. p. 1–4.

Markaki M, Stylianou Y. Voice pathology detection and discrimination based on modulation spectral features. IEEE Transactions on Audio, Speech, and Language Processing 2011;19(7):1938–48.

Martin GE, Klusek J, Estigarribia B, Roberts JE. Language characteristics of individuals with Down syndrome. Topics in Language Disorders 2009;29(2):112.

Martinez J, Perez H, Escamilla E, Suzuki MM. Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques. In: Electrical Communications and Computers (CONIELECOMP). IEEE; 2012. p. 248–51.

Martínez MH, Duran XP, Navarro JN. Attention deficit disorder with or without hyperactivity or impulsivity in children with Down's syndrome. International Medical Review on Down Syndrome 2011;15(2):18–22.

Moran MJ. Identification of Down's syndrome adults from prolonged vowel samples. Journal of communication disorders 1986;19(5):387–94.

Moran MJ, Gilbert HR. Selected acoustic characteristics and listener judgments of the voice of Down syndrome adults. American journal of mental deficiency 1982;.

Moura CP, Cunha LM, Vilarinho H, Cunha MJ, Freitas D, Palha M, Pueschel SM, Pais-Clemente M. Voice parameters in children with Down syndrome. Journal of Voice 2008;22(1):34–42.

Pentz Jr AL. Formant amplitude of children with Down syndrome. American journal of mental deficiency 1987;92(2):230–3.

Rochet-Capellan A, Dohen M. Acoustic characterisation of vowel production by young adults with Down syndrome. In: 18th International Congress of Phonetic Sciences (ICPhS 2015). 2015. .

Rodger R. Voice quality of children and young people with Down's Syndrome and its impact on listener judgement. Ph.D. thesis; Queen Margaret University; 2009.

Saz O, Simón J, Rodríguez W, Lleida E, Vaquero C, et al. Analysis of acoustic features in speakers with cognitive disorders and speech impairments. EURASIP Journal on Advances in Signal Processing 2009;2009:1.

Schiel F. Automatic phonetic transcription of non-prompted speech. In: International Congress of Phonetic Sciences (ICPhS). 1999. p. 607–10.

Schuller BW, Steidl S, Batliner A, Hirschberg J, Burgoon JK, Baird A, Elkins AC, Zhang Y, Coutinho E, Evanini K. The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. In: INTERSPEECH. 2016. p. 2001–5.

Seifpanahi S, Bakhtiar M, Salmalian T. Objective vocal parameters in Farsi-speaking adults with Down syndrome. Folia Phoniatrica et Logopaedica 2011;63(2):72–6.

Shott SR, Joseph A, Heithaus D. Hearing loss in children with Down syndrome. International journal of pediatric otorhinolaryngology 2001;61(3):199–205.

Van Borsel J, Vandermeulen A. Cluttering in Down syndrome. Folia Phoniatrica et Logopaedica 2008;60(6):312–7.

Wold DC MJ. Preliminary perceived voice deviations and hearing disorders of adults with Down's syndrome. Perceptual and Motor Skills 1979;49:564–564.

Wuang YP, Chiang CS, Su CY, Wang CC. Effectiveness of virtual reality using Wii gaming technology in children with Down syndrome. Research in developmental disabilities 2011;32(1):312–21.

Yuan J, Liberman M. Automatic measurement and comparison of vowel nasalization across languages. In: Proceedings of ICPhS. 2011. .

Zampini L, Fasolo M, Spinelli M, Zanchi P, Suttora C, Salerni N. Prosodic skills in children with Down syndrome and in typically developing children. International Journal of Language & Communication Disorders 2016;51(1):74–83.

Appendix A. Description of the features

The tables included in this appendix describe the features used in each of the domains. Frequency features are presented in Table A.9. Energy features are described in Table A.10. Temporal features are explained in Table A.11. Spectral features are presented in Tables A.12 and A.13.

Feature	Description
F0_stddevRisingSlope (F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope)	Standard deviation of the slope of rising signal parts of F0
jitter_stddevNorm (jitterLocal_sma3nz_stddevNorm)	Coefficient of variation of the deviations in individual consecutive F0 period lengths
jitter_mean (jitterLocal_sma3nz_amean)	Mean of the deviations in individual consecutive F0 period lengths
F0_pctlrange (F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2)	Range of 20-th to 80-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_percentile20 (F0semitoneFrom27.5Hz_sma3nz_percentile20.0)	Percentile 20-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_percentile50 (F0semitoneFrom27.5Hz_sma3nz_percentile50.0)	Percentile 50-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_mean (F0semitoneFrom27.5Hz_sma3nz_amean)	Mean of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_stddevNorm (F0semitoneFrom27.5Hz_sma3nz_stddevNorm)	Coefficient of variation of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
F0_percentile80 (F0semitoneFrom27.5Hz_sma3nz_percentile80.0)	Percentile 80-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz

Table A.9: Frequency features explained. All functionals are applied to voiced regions only. Text in brackets shows the original name of the eGeMAPS features

Feature	Description
loudness_percentile20 (loudness_sma3_percentile20.0)	Percentile 20-th of estimate of perceived signal intensity from an auditory spectrum
loudness_percentile50 (loudness_sma3_percentile50.0)	Percentile 50-th of estimate of perceived signal intensity from an auditory spectrum
loudness_mean (loudness_sma3_amean)	Mean of estimate of perceived signal intensity from an auditory spectrum
loudness_percentile80 (loudness_sma3_percentile80.0)	Percentile 80-th of estimate of perceived signal intensity from an auditory spectrum
loudness_pctlrange02 (loudness_sma3_pctlrange0-2)	Range of 20-th to 80-th of estimate of perceived signal intensity from an auditory spectrum
loudness_stddevRisingSlope (loudness_sma3_stddevRisingSlope)	Standard deviation of the slope of rising signal parts of loudness
loudness_stddevNorm (loudness_sma3_stddevNorm)	Coefficient of variation of estimate of perceived signal intensity from an auditory spectrum
shimmer_mean (shimmerLocaldB_sma3nz_amean)	Mean of difference of the peak amplitudes of consecutive F0 periods
shimmer_stddevNorm (shimmerLocaldB_sma3nz_stddevNorm)	Coefficient of variation of difference of the peak amplitudes of consecutive F0 periods

Table A.10: Energy features explained. All functionals are applied to voiced and unvoiced regions together. Text in brackets shows the original name of the eGeMAPS features

Feature	Description
silencePercentage	Duration percentage of unvoiced regions
silencesMean	Mean of unvoiced regions
StddevVoicedSegmentLengthSec	Standard deviation of continuously voiced regions
MeanUnvoicedSegmentLength	Mean of unvoiced regions
silencesPerSecond	The number of silences per second
VoicedSegmentsPerSec	The number of continuous voiced regions per second
loudnessPeaksPerSec	The number of the loudness peaks per second
MeanVoicedSegmentLengthSec	Mean of continuously voiced regions
soundingPercentage	Duration percentage of voiced regions

Table A.11: Temporal features explained

Feature	**Description**
mfcc3_stddevNorm (mfcc3_sma3_stddevNorm)	Coefficient of variation of Mel-Frequency Cepstral Coefficient 3
slopeV0500_mean (slopeV0-500_sma3nz_amean)	Mean of linear regression slope of the logarithmic power spectrum within 0-500 Hz band in voiced regions
mfcc2V_mean (mfcc2V_sma3nz_amean)	Mean of Mel-Frequency Cepstral Coefficient 2 in voiced regions
mfcc4_stddevNorm (mfcc4_sma3_stddevNorm)	Coefficient of variation of Mel-Frequency Cepstral Coefficient 4
slopeUV0500_mean (slopeUV0-500_sma3nz_amean)	Mean of linear regression slope of the logarithmic power spectrum within 0-500 Hz band in unvoiced regions
slopeV0500_stddevNorm (slopeV0-500_sma3nz_stddevNorm)	Coefficient of variation of linear regression slope of the logarithmic power spectrum within 0-500 Hz band in voiced regions
mfcc2_stddevNorm (mfcc2_sma3_stddevNorm)	Coefficient of variation of Mel-Frequency Cepstral Coefficient 2
mfcc2_mean (mfcc2_sma3_amean)	Mean of Mel-Frequency Cepstral Coefficient 2
alphaRatioUV_mean (alphaRatioUV_sma3nz_amean)	Mean of the ratio of the summed energy from 50-1000 Hz and 1-5 kHz in unvoiced regions
logRelF0H1A3_stddevNorm (logRelF0-H1-A3_sma3nz_stddevNorm)	Coefficient of variation of the ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3) in voiced regions
hammarbergIndexUV_mean (hammarbergIndexUV_sma3nz_amean)	Mean of the ratio of the strongest energy peak in the 0-2 kHz region to the strongest peak in the 2-5 kHz region in unvoiced regions
mfcc3V_stddevNorm (mfcc3V_sma3nz_stddevNorm)	Coefficient of variation of Mel-Frequency Cepstral Coefficient 3 in voiced regions
mfcc4V_stddevNorm (mfcc4V_sma3nz_stddevNorm)	Coefficient of variation of Mel-Frequency Cepstral Coefficient 4 in voiced regions
mfcc4_mean (mfcc4_sma3_amean)	Mean of Mel-Frequency Cepstral Coefficient 4
spectralFlux_mean (spectralFlux_sma3nz_amean)	Mean of the difference of the spectra of two consecutive frames
spectralFluxUV_mean (spectralFluxUV_sma3nz_amean)	Mean of the difference of the spectra of two consecutive frames in unvoiced regions
spectralFluxV_mean (spectralFluxV_sma3nz_amean)	Mean of the difference of the spectra of two consecutive frames in voiced regions

Table A.12: Spectral features explained (part1). If nothing is said, the features are applied to voiced and unvoiced regions together. Text in brackets shows the original name of the eGeMAPS features

Feature	Description
alphaRatioV_mean (alphaRatioV_sma3nz_amean)	Mean of the ratio of the summed energy from 50-1000 Hz and 1-5 kHz in voiced regions
mfcc4V_mean (mfcc4V_sma3nz_amean)	Mean of Mel-Frequency Cepstral Coefficient 4 in voiced regions
hammarbergIndexV_mean (hammarbergIndexV_sma3nz_amean)	Mean of the ratio of the strongest energy peak in the 0-2 kHz region to the strongest peak in the 2-5 kHz region in voiced regions
mfcc1V_mean (mfcc1V_sma3nz_amean)	Mean of Mel-Frequency Cepstral Coefficient 1 in voiced regions
hammarbergIndexV_stddevNorm (hammarbergIndexV_sma3nz_stddevNorm)	Coefficient of variation of the ratio of the strongest energy peak in the 0-2 kHz region to the strongest peak in the 2-5 kHz region in voiced regions
mfcc1_mean (mfcc1_sma3_amean)	Mean of Mel-Frequency Cepstral Coefficient 1
logRelF0H1A3_mean (logRelF0-H1-A3_sma3nz_amean)	Mean of the ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3) in voiced regions
F3amplitudeLogRelF0_mean (F3amplitudeLogRelF0_sma3nz_amean)	Mean of the ratio of the energy of the spectral harmonic peak at the third formant´s centre frequency to the energy of the spectral peak at F0 in voiced regions
F3amplitudeLogRelF0_stddevNorm (F3amplitudeLogRelF0_sma3nz_stddevNorm)	Coefficient of variation of the ratio of the energy of the spectral harmonic peak at the third formant´s centre frequency to the energy of the spectral peak at F0 in voiced regions
slopeV5001500_mean (slopeV500-1500_sma3nz_amean)	Mean of linear regression slope of the logarithmic power spectrum within 500-1500 Hz band in voiced regions
F2amplitudeLogRelF0_mean (F2amplitudeLogRelF0_sma3nz_amean)	Mean of the ratio of the energy of the spectral harmonic peak at the second formant´s centre frequency to the energy of the spectral peak at F0 in voiced regions
F2amplitudeLogRelF0_stddevNorm (F2amplitudeLogRelF0_sma3nz_stddevNorm)	Coefficient of variation of the ratio of the energy of the spectral harmonic peak at the second formant´s centre frequency to the energy of the spectral peak at F0 in voiced regions
F1bandwidth_stddevNorm (F1bandwidth_sma3nz_stddevNorm)	Coefficient of variation of the bandwidth of first formant in voiced regions
F1frequency_stddevNorm (F1frequency_sma3nz_stddevNorm)	Coefficient of variation of the centre frequency of first formant in voiced regions
F3frequency_stddevNorm (F3frequency_sma3nz_stddevNorm)	Coefficient of variation of the centre frequency of third formant in voiced regions
spectralFlux_stddevNorm (spectralFlux_sma3_stddevNorm)	Coefficient of variation of the difference of the spectra of two consecutive frames
F3frequency_mean (F3frequency_sma3nz_amean)	Mean of the centre frequency of third formant in voiced regions

Table A.13: Spectral features explained (part2). If nothing is said, the features are applied to voiced and unvoiced regions together. Text in brackets shows the original name of the eGeMAPS features