Universidad de Valladolid

DOCTORAL PROGRAM IN COMPUTER SCIENCE

DOCTORAL THESIS:

DESIGN AND EVALUATION OF MOBILE COMPUTER-ASSISTED PRONUNCIATION TRAINING TOOLS FOR SECOND LANGUAGE LEARNING

Thesis submitted by Cristian TEJEDOR-GARCÍA as part of the requirements for the degree of DOCTOR OF COMPUTER SCIENCE with mention of INTERNATIONAL DOCTORATE by the University of Valladolid

Supervised by:

Dr. Valentín CARDEÑOSO-PAYO
Dr. David ESCUDERO-MANCEBO

Valladolid, 2020

Universidad de Valladolid

PROGRAMA DE DOCTORADO EN INFORMÁTICA

TESIS DOCTORAL:

DISEÑO Y EVALUACIÓN DE HERRAMIENTAS MÓVILES PARA EL ENTRENAMIENTO ASISTIDO POR ORDENADOR DE LA PRONUNICACIÓN PARA EL APRENDIZAJE DE IDIOMAS

Presentada por Cristian TEJEDOR-GARCÍA para optar al grado de DOCTOR EN INFORMÁTICA con MENCIÓN DE DOCTORADO INTERNACIONAL por la Universidad de Valladolid

Dirigida por:

Dr. Valentín CARDEÑOSO-PAYO
Dr. David ESCUDERO-MANCEBO

Valladolid, 2020

Abstract

The quality of speech technology (automatic speech recognition, ASR, and text-to-speech, TTS) has considerably improved and, consequently, an increasing number of computer-assisted pronunciation (CAPT) tools has included it. However, pronunciation is one area of teaching that has not been developed enough since there is scarce empirical evidence assessing the effectiveness of tools and games that include speech technology in the field of pronunciation training and teaching. This PhD thesis addresses the design and validation of an innovative CAPT system for smart devices for training second language (L2) pronunciation. Particularly, it aims to improve learner's L2 pronunciation at the segmental level with a specific set of methodological choices, such as learner's first and second language connection (L1–L2), minimal pairs, a training cycle of exposure–perception–production, individualistic and social approaches, and the inclusion of ASR and TTS technology. 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. Data is automatically gathered by the CAPT systems to give an immediate specific feedback to users and to analyze all results. The protocols, metrics, algorithms, and methods necessary to statistically analyze and discuss the results are also detailed. The two main L2 tested during the experimental procedure are American English and Spanish. The different CAPT prototypes designed and validated in this thesis, and the methodological choices that they implement, allow to accurately measuring the relative pronunciation improvement of the individuals who trained with them. Both rater's subjective scores and CAPT's objective scores show a strong correlation, being useful in the future to be able to assess a large amount of data and reducing human costs. Results also show an intensive practice supported by a significant number of activities carried out. In the case of the controlled experiments, students who worked with the CAPT tool achieved better pronunciation improvement values than their peers in the traditional in-classroom instruction group. In the case of the challenge-based CAPT learning game proposed, the most active players in the competition kept on playing until the end and achieved significant pronunciation improvement results.

***Keywords*:** 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.

Resumen

El aumento de la mejora de la calidad de las tecnologías del habla (reconocimiento automático y síntesis del habla, ASR y TTS, respectivamente) trae consigo el incremento del número de herramientas para el entrenamiento de la pronunciación asistida por ordenador (CAPT). Sin embargo el uso de la tecnología para el entrenamiento de la pronunciación no está aún extendido de forma masiva debido a, entre otras causas, la falta de evidencia empírica sobre la eficacia de las aplicaciones que incluyen tecnología del habla para la enseñanza y entrenamiento de la pronunciación. Esta tesis doctoral aborda el diseño y la validación de un innovador sistema CAPT para dispositivos inteligentes para el entrenamiento de la pronunciación de lengua extranjera (L2). En concreto, tiene como objetivo mejorar la pronunciación L2 a nivel segmental del alumno mediante un conjunto específico de opciones metodológicas de carácter individual y social, como la conexión entre la lengua materna (L1) y la L2, un ciclo de entrenamiento de exposición–percepción–producción con pares mínimos, y la inclusión de tecnología ASR y TTS. La investigación experimental realizada aplicando estas opciones metodológicas con usuarios reales valida la eficacia de los prototipos CAPT desarrollados para los cuatro experimentos principales de esta tesis. Dichos sistemas CAPT recogen y analizan la información de interacción del usuario para proporcionarle una retroalimentación específica e inmediata. Gracias a los protocolos, métricas, algoritmos y métodos descritos en este trabajo, los resultados se analizan estadísticamente y discuten. Los dos principales L2 probados durante el procedimiento experimental son el inglés americano y español. Los diferentes prototipos CAPT diseñados y validados en esta tesis, y las opciones metodológicas que implementan, permiten medir con precisión la mejora de la pronunciación relativa de los estudiantes que entrenaron con ellos. Tanto las puntuaciones subjetivas de los evaluadores, como las objetivas de los sistemas CAPT, muestran una alta correlación, siendo estas últimas útiles en el futuro para la evaluación de una gran cantidad de datos y la reducción de tareas para los evaluadores. El número significativo de actividades llevadas a cabo por los participantes respalda una práctica intensa en la experimentación. En el caso de los experimentos controlados, los estudiantes que trabajaron con la herramienta CAPT lograron mayores valores de mejora de la pronunciación que sus compañeros en el grupo de aprendizaje tradicional en el aula. En el caso del juego educativo CAPT basado en desafíos, los jugadores más activos en la competición lograron resultados de mejora de pronunciación significativos.

***Palabras clave*:** Entrenamiento de la pronunciación asistida por ordenador (CAPT), pronunciación de segunda lengua (L2), reconocimiento automático del habla (ASR), síntesis de habla (TTS), aprendizaje autónomo, entornos educativos, juego educativo móvil, pares mínimos.

Acknowledgements

In the following paragraphs I would like to thank the people who have accompanied and supported me in this exciting and intensive chapter of my life to obtain the PhD degree. It has not only scientifically educated me, but also developed myself personally, making me realize my (various) limitations and taking on new responsibilities.

This PhD thesis has been carried out in the ECA-SIMM research group with the Department of Computer Science of the University of Valladolid. My first contact with the research group was when I was finishing my master's degree in Computer Science at the middle of the year 2015. They were looking for one candidate to start researching in foreign pronunciation teaching with speech technology and mobile applications within a research project. I was really interested in what I was doing from the very first day in the laboratory.

First of all, I am deeply indebted to my two supervisors for guiding me throughout the process of becoming a PhD graduate. Dr. Valentín Cardeñoso-Payo, there are not enough words to describe your contribution to my life during these last five years. We have shared many trips, anecdotes, and hard days of work. Among your innumerable virtues I would like to point out your capacity to synthesize knowledge and for finding solutions to almost every problem. I would also like to extend my deepest gratitude to Dr. David Escudero-Mancebo. Your proactive personality and your unique ability for finding significant results have helped me to reach the objectives of this thesis. You are such a nice person who has not only earned a successful research career, but is well-aware of the environmental issues.

I am also extremely grateful to my colleague Dr. César González-Ferreras. I cannot thank you enough for everything that you have done for making this work a possible one. I really appreciate the pragmatic and helpful way that you have proceeded whenever I needed help.

I would also like to take this opportunity to thank Dr. Enrique Cámara-Arenas for sharing his ideas and helpful suggestions for this work. Your love for writing and endless creativity have been essential to this thesis to succeed.

I want to thank Dr. Mario Corrales-Astorgano, my colleague from the ECA-SIMM group. Sharing my time and experiences with you has been very rewarding. I am also grateful for your selfless help. Although we started and almost finished the predoctoral program at the same time, our research careers have just started.

I also want to thank Dr. María Jesús Rodríguez-Triana, Dr. Tobias Ley, and Dr. Luis Pablo Prieto Santos for giving me the opportunity of being part of your research group during three months at the Centre of Excellence in Educational Innovation at the University of Tallinn, Estonia. I am so grateful for letting me into your international family and in what I consider my second home. Traveling to such a wonderful (and different) city is exciting but also a little overwhelming. Thanks also to Mr.

Pankaj Chejara, Mrs. Ana Sofia Chermont, Mr. Shashi Kant, Mrs. Milena Sarmiento, and Mr. Gerti Pishtari. I never thought I would meet such amazing people. You have made it easier to be there. I am pretty sure we can continue collaborating together. *Aitäh!*

The experience at the Servei de Tractament de la Parla i del So of the Autonomous University of Barcelona during one month in the summer of 2017 was not only a profitable but an enriching adventure. *Moltes gràcies pel vostre gran suport*, contribution, and exchange of knowledge, Dr. María J. Machuca and Dr. Antonio Ríos.

Tänan väga! Mrs. Katrin Leppik. It was a pleasure to meet you. Your dedication to our collaboration in one of the experimental prototypes is much appreciated. I firmly believe my trip to Tartu would not have been the amazing experience it was, if you and Rene had not allowed me to share those incredible days together. I hope we can continue collaborating and I wish you the best for your promising career.

I am also so thankful for the useful comments and suggestions in the reviewing phase of this manuscript given by Dr. Laura Colantoni of the University of Toronto and Dr. Antonio Origlia of the University of Naples Federico II.

I really appreciate the charming welcome in the Department of Computer Science. Teaching at the university level has been another positive challenge during this period. *Doy las gracias en especial a* Dr. Belarmino Pulido, Mrs. Alma María Pisabarro, Dr. Yania Crespo, Dr. Pablo de la Fuente, Dr. Alejandra Martínez-Monés, Mr. César Vaca, and Dr. Carlos Enrique Vivaracho.

During these years I have attended several conferences and workshops where I established new contacts and friends who I would like to thank for their support and contributions to the experimental work: Dr. Valle Flores-Lucas, Mr. Takuya Kimura, Dr. Andreia Schurt Rauber, Dr. Anabela Rato, Dr. Junming Yao, Dr. Cristiane Silva, Dr. Andrea Cesco, Mr. Hernán Camilo Urón, Mrs. Hilola Ruziyeva, Mrs. Carolina Regidor, and Mrs. Amaia Olmo.

I would also like to thank the rest of researchers who collaborated in any way with the ECA-SIMM research group and other groups of the Department, and partially contributed to this thesis (Mrs. Verónica Barroso, Mr. Eduardo Gutiérrez, Mr. Diego Cubero, Mr. Rafael Sillero, Mr. Mario Cartón, Mr. Ismael Taboada, and all the students of final degree projects who worked with us in the group).

I truly appreciate the immense (and tedious) work done by the professional members of the *Teatro Pie Izquierdo* of Valladolid by recording thousands of word utterances. The Language Learning Center of the University of Valladolid has been always our first place to find students for the experiments. Most of the work of this thesis could not have done without its support. Furthermore, special thanks to *IdiomApps*, *La Casa del Español*, and *Warwick House* for helping us to test our prototypes. Thank you Pablo Villanueva, Gonzalo Bajeneta, and the rest members of the *Fundación General Universidad de Valladolid* who helped us with the intellectual property register of the software.

Finally, I want to thank the University of Valladolid for providing me with the financial support necessary to carry out this work (predoctoral research fellowship, 2015, and predoctoral short-term fellowship, 2019 —budgetary application: 180113-541A.2.01-691). Thanks a lot for making it possible.

Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

3-D	(**3**)Three **D**imensional
App	**A**pplication (software)
ASR	**A**utomatic **S**peech **R**ecognition
AVR	**A**utomatic **V**oice **R**ecognition
CALL	**C**omputer-**A**ided **L**anguage **L**earning
CAPT	**C**omputer-**A**ssisted **P**ronunciation **T**raining
CEFR	**C**ommon **E**uropean **F**ramework **R**eference
CF	**C**orrective **F**edback
CMVN	**C**epstral **M**ean and **V**ariance **N**ormalization
CN	**C**hinese
cn_ZH	Simplified **C**hinese (Mainland China)
COP	**C**lash **o**f **P**ronunciations (prototype)
CPU	**C**entral **P**rocessing **U**nit
DE	German
de_DE	German (Germany)
DNN	**D**eep **N**eural **N**etworks
ECA-SIMM	**E**ntornos de **C**omputación **A**vanzada y **S**istemas de **I**nteracción **M**ulti**m**odal
EFL	**E**nglish as **F**oreign **L**anguage
EME-E	**E**scala de **M**otivación **E**ducativa (**E**spaña)
EN	**E**nglish
en_US	American **E**nglish
EÑ	**E**stoñol (prototype)
ES	**S**panish
es_ES	Castilian **S**panish
ET	**E**stonian
et_EE	**E**stonian (**E**stonia)
EVow	**E**nglish **V**owels (prototype)
fMLLR	**F**eature **S**pace **M**aximum **L**ikelihood **L**inear **R**egression
FST	**F**inite-**S**tate **T**ransducers
GCSTT	**G**oogle **C**loud **S**peech-**T**o-**T**ext
GMM	**G**aussian **M**ixture **M**odel
GOP	**G**oodness **o**f **P**ronunciation
GUI	**G**raphical **U**ser **I**nterface
HMM	**H**idden **M**arkov **M**odel
ICT	**I**nformation and **C**ommunications **T**echnology
IEEE	**I**nstitute of **E**lectrical and **E**lectronics **E**ngineers
IPA	**I**nternational **P**honetic **A**lphabet
JCR	**J**ournal **C**itation **R**eport
JÑ	**J**apañol (prototype)
JP	**J**apanese

jp_JP	**J**apanese (**J**apan)
JSON	**J**ava**S**cript **O**bject **N**otation
L1	First(**1**) **L**anguage
L2	Second/Foreign(**2**) **L**anguage
LDA	**L**inear **D**iscriminat **A**nalysis
LDC	**L**inguistic **D**ata **C**onsortium
LL	**L**anguage **L**earning
MALL	**M**obile-**A**ssisted **L**anguage **L**earning
MFCC	**M**el **F**requency **C**epstral **C**oefficients
MP	**M**inimal **P**airs (prototype)
NCM	**N**ative **C**ardinality **M**ethod
OOV	**O**ut-**o**f-**v**ocabulary (context)
OS	**O**perating **S**ystem
PC	**P**ersonal **C**omputer
PhD	**D**octor **o**f **P**hilosophy
PLP	**P**erceptual **L**inear **P**rediction
PPV	**P**ositive **P**redictive **V**alue
PT	**P**or**t**uguese
pt_BR	Brazilian **P**or**t**uguese
pt_PT	European **P**or**t**uguese
RO	**R**esearch **O**bjective
RQ	**R**esearch **Q**uestion
SAMPA	**S**peech **A**ssessment **M**ethods **P**honetic **A**lphabet
SAT	**S**peaker **A**daptive **T**raining
SCI	**S**ocial **C**itation **I**ndex
SGMM	**S**ubspace **G**aussian **M**ixture **M**odel
SLA	**S**econd **L**anguage **A**cquisition
SVN	Apache **S**ub**v**ersio**n**
TPR	**T**rue-**P**ositive-**R**ecall
TTS	**T**ext-**T**o-**S**peech
TTT	**T**ip**T**op**T**alk! (prototype)
UK	**U**nited **K**ingdom
USA	**U**nited **S**tates of **A**merica
UX	**U**ser E**x**perience
VTLN	**V**ocal **T**ract **L**ength **N**ormalization
vs.	**V**er**s**us
WER	**W**ord **E**rror **R**ate
WFSTs	**W**eighted **F**inite **S**tate **T**ransducers

To my beloved Teresa, Bernardo
and Raúl. You were, are, and will
always be by my side.

Capítulo R1

Resumen de la Tesis Doctoral

El éxito de la comunicación en un idioma extranjero depende en gran medida de la inteligibilidad, comprensión, acento y fluidez del habla. Los cursos de aprendizaje de idiomas se han centrado tradicionalmente, sin embargo, en otras áreas de habilidades lingüísticas, como la gramática o la comprensión escrita. Por un lado, los nuevos enfoques que ofrecen las recientes innovaciones en las tecnologías del habla mejoran significativamente el rendimiento del reconocimiento y síntesis del habla. Dicha tecnología se puede integrar en sistemas pedagógicos para dispositivos inteligentes actuales para el entrenamiento de la pronunciación mediante aplicaciones que complementen el aprendizaje. Esto permite a los estudiantes usarlas de forma continua y autónoma. Por otro lado, las aplicaciones de juegos educativos tienen un enorme potencial para la educación y, en particular, para el aprendizaje de idiomas. El proceso de aprendizaje se ve afectado por la participación social que implican dichos juegos, cuya utilidad y eficacia deben ser evaluadas. No obstante, existe escasa evidencia experimental de la eficacia del uso aplicaciones tecnológicas para el entrenamiento y mejora de la pronunciación extranjera.

Este capítulo plantea el contexto y motivación de este trabajo de tesis doctoral en relación a los temas mencionados anteriormente. Además, se describen de forma general las contribuciones aportadas en la tesis doctoral, al resolver las cuestiones planteadas en las preguntas de investigación, y conseguir los objetivos descritos.

R1.1 Motivación

La demanda actual de aprendizaje de segunda lengua (SLA) es muy alta. A finales de 2016 existían 912 millones de estudiantes de segunda lengua (L2) en todo el mundo [1], una séptima parte de la población mundial. Esto es, en cierta medida, por la necesidad de comunicación entre personas de cualquier lugar del mundo por medio de la tecnología actual que permite este proceso. No obstante, la gran cantidad y diferencias entre idiomas y culturas pueden ser una barrera para conseguir una comunicación exitosa.

Se estima que cada persona tiene acceso a alrededor de 6 dispositivos inteligentes en el año 2020 [1]. Dichos dispositivos forman parte de la tecnología educativa (e-learnig) y autoaprendizaje; alternativas interesantes a los cursos tradicionales en el aula. En particular, los sistemas de aprendizaje de idiomas asistido por ordenador (CALL) y por dispositivos móviles (MALL) integran tecnología avanzada muy atractiva para el aprendizaje de idiomas y que puede ayudar en el proceso de aprendizaje y enseñanza de manera eficiente. Sin llegar a reemplazar a los tutores humanos, pueden desempeñar un papel complementario en la educación al aumentar la eficiencia y la motivación del proceso de aprendizaje, se pueden utilizar en cualquier lugar, momento y tantas veces como se desee.

Actualmente, el número de juegos educativos para el aprendizaje de idiomas está aumentando, dado que la inclusión de elementos de juego en herramientas educativas favorece un mejor rendimiento individual [2]. Estudios recientes detallan que la motivación y el compromiso del alumno mejoran no solo dentro sino también fuera del aula [3]. Aunque la inclusión de elementos sociales y competitivos en cualquier sistema pedagógico debe hacerse con precaución, existen estudios que indican que la competitividad en el contexto del aprendizaje basado en juegos facilita el logro de objetivos educativos [4] y fomenta la cooperación como un elemento de apoyo al trabajo en clase [5].

Los sistemas para el entrenamiento de la pronunciación asistida por computador (CAPT) dan soporte a investigaciones y prácticas innovadoras que favorecen a la transformación del aprendizaje de idiomas, creando oportunidades para revisar las viejas ideas y desafiar las creencias establecidas [6]. CAPT es una subárea importante de CALL y MALL en constante cambio, que combina la retroalimentación correctiva y la evaluación automática de la calidad de la pronunciación, entre otras funcionalidades proporcionadas por las tecnologías del habla incorporadas. Las más comunes son el reconocimiento automático de voz (ASR) y la síntesis de habla (TTS), que transforman la voz en texto escrito, y viceversa, respectivamente. Hoy en día, estos sistemas están respaldados por una enorme cantidad de datos y algoritmos complejos que mejoran significativamente su calidad. Por ejemplo, Google reportó que sus recientes avances en el aprendizaje automático aplicado a TTS han ayudado a generar formas de onda de voz 1000 veces más rápido que antes (generar un segundo de audio solo tarda 50 milisegundos), y han logrado calificaciones más de un 20 % mejores que las voces estándar [7]. Además, en el campo del reconocimiento de voz, Google también ha alcanzado una tasa de precisión de palabras del 95 % para el idioma inglés, por lo tanto, alcanzando el umbral de precisión humana [8]. Es probable que se obtengan mejores tasas en el futuro cercano con el uso de técnicas de redes neuronales profundas (DNN) más sofisticadas, y mayores unidades de procesamiento central (CPU) [9], [10]. Aunque hasta 2014 se han reportado pocos estudios de investigación revisados por pares sobre CAPT (solo un 26.9 % de los 75 estudios del estado de la cuestión resumidos en [11]), los sistemas CAPT están evolucionando y apareciendo cada vez más debido a las mejoras y las nuevas posibilidades que ofrecen [12].

Una metodología de entrenamiento correcta debe abordar adecuadamente los aspectos de la retroalimentación automática instantánea y el diseño de actividades y elementos de enseñanza de acuerdo con la lengua materna (L1) y L2 del alumno, a fin de optimizar la eficacia de las herramientas CAPT y el tiempo de uso [13]. El proceso de aprendizaje para la adquisición de L2 se ve muy afectado por una percepción habitual bien establecida de los movimientos y sonidos articulatorios L1. A menudo conduce a errores e imprecisiones en la pronunciación L2 de los alumnos (es decir, una transferencia negativa del idioma [14]). En esta tesis se ha utilizado la técnica de pares mínimos, pares de palabras que varían en un solo sonido. El uso de pares mínimos puede aportar grandes beneficios en el aprendizaje y la enseñanza de la pronunciación, ya que aparecen en casi todos los idiomas y pueden contrastar los sonidos L1 y L2 [15].

En resumen, el aprendizaje de L2, y más precisamente, el entrenamiento de pronunciación de L2, está abierto a nuevos paradigmas de enseñanza. La posibilidad de que los alumnos entrenen en cualquier momento y en cualquier lugar, a su propio

ritmo, permite a los maestros proporcionar una instrucción individualizada en grupos pequeños en lugar de los tradicionales grupos de mayor tamaño. El hecho de que los estudiantes de hoy en día estén acostumbrados a la tecnología digital motiva el desarrollo de sistemas de aprendizaje CAPT para dispositivos inteligentes. Los alumnos están acostumbrados a utilizarlos para interactuar en entornos digitales para la comunicación, información, contacto social, reunión y análisis. Aunque pueden ser nativos digitales y estar cómodos e inmersos en la tecnología, dependen de maestros y expertos para aprender a través de los medios digitales. Además, las tecnologías ASR y TTS han mejorado drásticamente su rendimiento en los últimos años, pudiendo integrarse en los recursos educativos. Por lo tanto, el desafío actual es diseñar y adaptar cuidadosamente un sistema CAPT efectivo con no solo dicha tecnología, sino también con una metodología de entrenamiento, una evaluación de mejora de la pronunciación y una estrategia de retroalimentación correctiva, de acuerdo con las L1 y L2 del alumno.

R1.2 El Problema

Proporcionar un conjunto adecuado de actividades de entrenamiento para la pronunciación no es una tarea fácil ya que hay varios factores a tener en cuenta:

- **L1 del alumno**. Las similitudes y diferencias entre la primera y segunda lengua del estudiante varían la dificultad del proceso de aprendizaje.
- **El conjunto de actividades de entrenamiento personalizadas**. Dependiendo del nivel de pronunciación L2 del alumno y su desempeño, las actividades recomendadas deben ser individualizadas y adaptadas para que sean efectivas.
- **Evaluación de los resultados del alumno**. Se pueden proporcionar valoraciones subjetivas y objetivas a los usuarios, no solo al final de la experimentación, sino también durante el entrenamiento.
- **La retroalimentación proporcionada a los estudiantes**. Se necesita más retroalimentación de la que un maestro puede ofrecer en clase. Sin embargo, en algunos casos esta retroalimentación es insuficiente o demasiado difícil de entender para los alumnos.
- **La tecnología incluida en la metodología de entrenamiento** debe seleccionarse cuidadosamente ya que las puntuaciones de evaluación deben ser lo más precisas posible y orientarse al objetivo del entrenamiento.
- **Elementos motivacionales**, como los elementos de juego o la interacción con otros alumnos pueden influir en los resultados del entrenamiento, desviando a los estudiantes del objetivo real de mejora de la pronunciación o incluso desanimándolos.

Un trabajo de investigación que plantee los problemas descritos anteriormente debe abordarse desde una perspectiva multidisciplinar que incluya: metodología educativa, diseño de herramientas educativas, modelado de datos y técnicas de evaluación, entre otros.

Finalmente, es necesario establecer protocolos adecuados para recopilar y analizar los resultados de los experimentos, ya que la eficacia de la herramienta CAPT está influenciada por su escalabilidad y rendimiento. En primer lugar, la posibilidad de ampliar el conjunto de idiomas conduce a generalizar los conceptos y estrategias de entrenamiento, y a seleccionar correctamente la tecnología de voz necesaria. En

segundo lugar, la interacción del usuario con un sistema CAPT tiende a ser masiva en términos de datos de voz y de actividad registrada. En algunos casos es necesaria una importante inversión de dinero en dispositivos y servidores.

R1.3 Objetivos y Preguntas de Investigación

De acuerdo con los problemas encontrados y descritos en la sección anterior, el objetivo principal de esta tesis se define como:

Diseño y evaluación de una herramienta CAPT para dispositivos inteligentes que incorpore **tecnología TTS y ASR actual**; que ayude a los estudiantes a **trabajar de manera autónoma**, a su **propio ritmo**, y con la posibilidad de proveer **realimentación en tiempo real**.

Este objetivo principal se divide en cuatro objetivos de investigación específicos:

- **RO1**. Análisis y definición un conjunto de actividades, protocolos y elementos motivadores para la mejora de la pronunciación L2 con un sistema CAPT que integre tecnología TTS y ASR.
- **RO2**. Selección de las métricas más apropiadas para la evaluación del nivel de pronunciación del hablante.
- **RO3**. Diseño de un método semiautomático supervisado por expertos para la obtención de un conjunto específico de pares mínimos adaptados a los problemas de pronunciación L2, teniendo en cuenta la L1 del hablante y las limitaciones de la tecnología TTS y ASR.
- **RO4**. Selección y diseño de un sistema CAPT con tecnología TTS y ASR actual que proporcione una retroalimentación individualizada al hablante para mejorar la pronunciación en L2.

Para llevar a cabo el procedimiento experimental de esta tesis, **tres preguntas de investigación** (*RQs*) junto a sus subcuestiones (*Issues*) se identifican para validar los objetivos de investigación, categorizados por temas.

El primer tema está relacionado con la *factibilidad de integración de tecnología de voz actual (sistemas TTS y ASR) en herramientas CAPT*:

- **RQ1**. ¿Pueden los actuales sistemas TTS y ASR ser integrados con éxito y de una manera no obstrusiva[1] en la herramienta CAPT desarrollada?
  - ***Issue 1.1***. ¿Pueden los actuales sistemas TTS y ASR ayudar en la evaluación de diferentes grupos de hablantes según su nivel de pronunciación en la herramienta CAPT desarrollada?

El segundo tema se refiere a las *implicaciones de la metodología de entrenamiento* con las herramientas CAPT en la mejora de la pronunciación del alumno:

- **RQ2**. ¿En qué medida los aspectos de diseño metodológicos, como el uso de ejercicios basados en pares mínimos dentro del ciclo de actividades de entrenamiento propuesto en la herramienta CAPT desarrollada, afectan la mejora de la pronunciación del usuario?
  - ***Issue 2.1***. ¿Se puede medir una mejora relativa en la pronunciación del estudiante después de usar la herramienta CAPT?

[1] Del inglés: *obstructive*: que bloquea o impide la realización o consecución de un fin.

  - ***Issue 2.2***. Si existe dicha mejora, ¿se pueden obtener evidencias desde el punto de vista cuantitativo?
  - ***Issue 2.3***. ¿La herramienta revela cuáles son las dificultades reales de los usuarios (los sonidos más difíciles y las actividades de entrenamiento más difíciles)?

Finalmente, la última pregunta de investigación tiene como objetivo responder a *cómo los elementos de juego y los enfoques sociales afectan la implicación del alumno* en el entrenamiento de pronunciación con herramientas CAPT:

- **RQ3**. ¿En qué medida pueden afectar a la motivación, el rendimiento y el aprendizaje del usuario las versiones *gamificadas* de la herramienta?

R1.4 Metodología de Investigación

Se ha llevado a cabo una **metodología experimental** [16] durante todo el proceso de experimentación con un grupo multidisciplinar de investigadores y expertos para cumplir los objetivos y dar respuesta las preguntas de investigación propuestas en la tesis. En la metodología se definen 5 fases en cada iteración experimental:

1. **Identificación del problema de investigación**. El proceso comienza identificando claramente los problemas que se abordarán durante la investigación, comenzando con el análisis de las soluciones existentes en el estado de la cuestión y considerando qué posibles métodos llevarán a la solución.
2. **Planificación del estudio de investigación experimental**. Se diseña cuidadosamente el experimento para evaluar los objetivos y las preguntas de la investigación.
   (a) **Selección de participantes**. Se define la población objetivo, las reglas de participación, el tamaño de la muestra y los grupos.
   (b) **Variables**. Se definen diferentes métricas para medir las variables de investigación a partir de los resultados de los datos recopilados a través de los instrumentos.
   (c) **Protocolo de evaluación**. Las variables de investigación se miden antes, durante y después de realizar las actividades de entrenamiento.
   (d) **Desarrollo de herramienta CAPT**. Para cada experimento en esta tesis se desarrolla una innovadora herramienta CAPT.
3. **Puesta en marcha del experimento**. Al principio se forman los grupos de participantes. Después, cada usuario realiza las actividades de su grupo definidas en la fase anterior, y los datos experimentales relacionados con las variables del estudio y se recopilan los datos experimentales con instrumentos específicos para cada experimento.
4. **Análisis de datos**. Se analizan los datos recopilados, indicando qué indicadores son los relevantes para corroborar el éxito del experimento.
5. **Publicación de resultados**. Los resultados más relevantes se comparten en publicaciones tipo revistas y conferencias científicas mediante artículos, resúmenes, demostraciones y presentaciones.

R1.5 Estructura del Documento de Tesis

Este documento está estructurado en seis partes. En la **primera** (Capítulo 1), se presenta y motiva el tema principal de esta tesis, junto a los objetivos y preguntas de investigación, dando una visión global de la metodología de investigación realizada en este trabajo. En la **segunda parte**, se presenta y discute una revisión profunda del trabajo relacionado del estado de la cuestión respecto a las características principales de los sistemas y experimentos relacionados con los objetivos de esta tesis doctoral. En particular, en el Capítulo 2 se revisan las actividades para el entrenamiento de la pronunciación tradicionales y aplicadas a sistemas CAPT, junto a la posible integración de sistemas ASR y TTS en dichos sistemas. En el Capítulo 3 se discuten las estrategias del estado de la cuestión acerca de la evaluación de la mejora de la pronunciación para CAPT. En el Capítulo 4 se examinan las estrategias de retroalimentación correctiva adoptadas por los estudios CAPT en la literatura. En el Capítulo 5 se describen los fundamentos del aprendizaje individual y social aplicado al entrenamiento de la pronunciación. También se menciona el efecto de los elementos de *gamificación* en contextos de aprendizaje. Los detalles específicos sobre los conceptos, estrategias y elementos del marco experimental necesarios para la experimentación se especifican en la **tercera parte** de este documento. En particular, en el Capítulo 6 se incluye una revisión exhaustiva de las dimensiones comunes del procedimiento experimental en relación con el estado de la cuestión. En el Capítulo 7 se detalla cada experimento en profundidad, dando una visión evolutiva del trabajo realizado a lo largo de esta tesis. En el Capítulo 8 se discuten los resultados obtenidos en todos los experimentos para dar respuesta a las preguntas de investigación de esta tesis. En la **cuarta parte** (Capítulo 9) se resumen las conclusiones y se definen algunas direcciones futuras de este trabajo de tesis. Además, se enumeran las publicaciones, la financiación de la investigación, los logros y las atribuciones obtenidas durante el transcurso de esta tesis.

Los apéndices se incluyen en la **quinta parte** de este documento. En el Apéndice A se explica en profundidad el algoritmo para elaborar listas de pares mínimos diseñado en esta tesis. En el Apéndice B se presenta una visión general sobre las similitudes y diferencias de todos los experimentos de esta tesis a través de tablas comparativas. El Apéndice C muestra la lista de palabras de la prueba previa y posterior al entrenamiento de los experimentos correspondientes. El Apéndice D representa las características principales de los datos del corpus del habla reunidos en dos experimentos de esta tesis. El Apéndice E presenta la estructura de proyecto estándar de Kaldi y una lista de pasos para desarrollar un sistema ASR. Finalmente, en la **sexta** y última parte de esta tesis se incluyen las referencias bibliográficas siguiendo el formato de la Guía de Referencia de IEEE[2].

R1.6 Síntesis de Resultados y Contribuciones

En esta sección se muestra la visión de conjunto logrado en este trabajo, presentando los experimentos y sus resultados de forma resumida y justificada. También se señalan las publicaciones en las que se presenta más detalladamente cada uno de los experimentos.

En esta tesis se ha seguido un enfoque incremental y evolutivo en el diseño de cuatro experimentos, como se ha descrito en la Sección R1.4. Cada uno de ellos ha

[2] http://ieeeauthorcenter.ieee.org/wp-content/uploads/IEEE-Reference-Guide.pdf

sido cambiado y refinado como consecuencia del análisis de los resultados del anterior. Además, este trabajo ha formado parte de tres proyectos de investigación diferentes con equipos multidisciplinares de investigadores, cuyos objetivos también han influido en los pasos dados y el flujo de experimentación de esta tesis (véase la Sección 9.3.9 acerca de los detalles sobre la financiación).

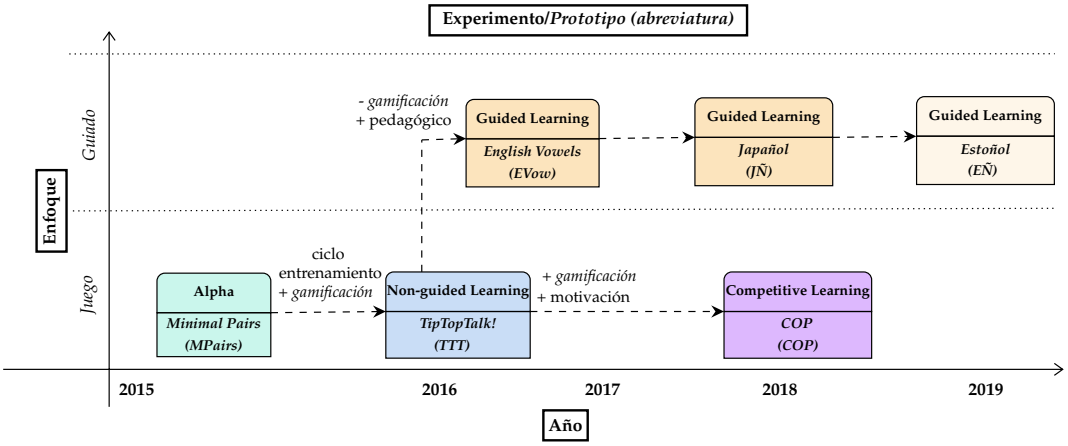

FIGURA R1.1: Diagrama de experimentos y prototipos.

Se han desarrollado cinco prototipos para los cuatro experimentos centrales de esta tesis (y uno más está en fase de experimentación) como se muestra en la Figura R1.1. La evolución de estos prototipos está alineada con dos enfoques principales: el primero (seguido por los tres experimentos en la parte inferior de la figura) propone incorporar elementos de *gamificación* y estrategias sociales; mientras que el segundo enfoque (los tres prototipos experimentales de la parte superior de la figura) sigue una aproximación de carácter individual mediante instrucciones de entrenamiento guiadas. A continuación describimos cada uno de los cuatro experimentos.

R1.6.1 Prototipo *Minimal Pairs*

El primer prototipo experimental de esta tesis (***Minimal Pairs*** del experimento ***Alpha***) fue una prueba de concepto para evaluar la posibilidad de incluir tecnología del habla de vanguardia (ASR y TTS) en un protocolo para el entrenamiento de la pronunciación. En particular, tres grupos de sujetos reales con diferentes niveles de habilidad L2 probaron dos sistemas de tecnología del habla de propósito general (Google ASR y Google TTS) con ejercicios de producción aislados con pares mínimos de palabras de inglés específicamente seleccionados por un experto. Estas palabras fueron incluidas en un sistema CAPT móvil implementado desde cero en el que todos los datos de interacción del usuario se recopilaron automáticamente. La posición de la palabra objetivo en la lista de hipótesis ASR junto a una puntuación sirvieron como métricas para relacionar el resultado final con el nivel de habilidad declarado por los hablantes. El uso aislado y voluntario del TTS fue mucho mayor por parte de los usuarios no nativos, aunque no se obtuvieron mejoras significativas tras su uso. Además, se encontraron limitaciones en la tecnología, ya que los hablantes nativos no completaron con éxito todas las actividades de pronunciación. Finalmente, las opiniones en la sesión de grupo focal sobre la herramienta CAPT sugirieron nuevas actividades, técnicas de retroalimentación y elementos motivacionales a tomar en cuenta para los próximos experimentos.

En este experimento comenzamos a arrojar luz sobre la pregunta de investigacion RQ1 (e *Issue 1.1*), tratando aspectos relativos a los objetivos RO1, RO2 y RO3. Los principales resultados han sido publicados en [17]. En la Sección 7.2 se encuentra la descripción completa de dicho experimento.

R1.6.2 Prototipo *TipTopTalk!*

Una vez comprobado el potencial y las limitaciones de la tecnología del habla en un sistema CAPT en el primer prototipo experimental, se diseñó un segundo prototipo llamado ***TipTopTalk!***, incluido en el segundo experimento (***Non-guided Learning***). Su objetivo era evaluar la mejora de la pronunciación de los posibles alumnos a lo largo de un tiempo prolongado (un mes) en el contexto de una competición en la que cada participante entrenaba a su propio ritmo, eligiendo las actividades que quisiera y viendo sus resultados reflejados en una clasificación común. Además de incluir elementos de juego, teniendo en cuenta las sugerencias previas de los usuarios, se incluyó el ciclo de actividades exposición–discriminación–producción para ampliar la variedad y calidad de entrenamiento. Finalmente, además del inglés, se incluyeron otros idiomas como el chino y español, para comprobar la posible integración de más idiomas en el mismo sistema CAPT. Las actividades de discriminación fueron las más entrenadas, lo que condujo a una mejora general de los participantes en dicha habilidad. Sin embargo, a pesar de la introducción de elementos de *gamificación* y la mejora mencionada en la discriminación, se detectó un estancamiento en la intensidad del entrenamiento y la mejora de la producción, siendo mayor esta pérdida de interés en los mejores jugadores. Los participantes también tendieron a entrenar los ejercicios de sonidos más fáciles para obtener resultados positivos.

Este experimento fue el primero en intentar responder a las preguntas de investigación RQ2 (e *Issue 2.1*, *Issue 2.2*, *Issue 2.3*) y RQ3, junto a la pregunta de investigacion RQ1 (e *Issue 1.1*), tratando aspectos relativos a los objetivos RO1, RO2, RO3 y RO4. Los principales resultados han sido publicados en [18], [19], [20], [21], [22]. En la Sección 7.3 se encuentra la descripción completa de dicho experimento.

R1.6.3 Prototipos *English Vowels*, *Japañol* y *Estoñol*

Tanto el estancamiento en la mejora de la producción, como el descenso del número de actividades realizadas por los mejores participantes, siendo éstas mayoritariamente las más sencilllas, motivaron la investigación de nuevo enfoque de entrenamiento, más pedagógico, guiado e individualizado (experimento ***Guided Learning***). El objetivo principal era entrenar la pronunciación del usuario guiándolo a través de un sistema CAPT con una retroalimentación personalizada y más precisa, basada en los resultados que iba obteniendo el alumno. Se desarrollaron dos prototipos para el tercer experimento, llamados ***English Vowels*** para nativos españoles cuya L2 es el inglés y ***Japañol*** para nativos japoneses cuya L2 es el español, Además otro prototipo está siendo llevado a cabo al final del período doctoral en colaboración con la Universidad de Tartu, Estonia, llamado *Estoñol*, que sigue la misma filosofía de trabajo que los dos anteriores, pero para hablantes de español que pretenden mejorar su pronunciación estonia. En estos prototipos se siguió una estrategia Pre/Post para determinar la mejora del nivel de pronunciación de los participantes de diferentes grupos de entrenamiento (con aplicación CAPT y en el aula) durante unas sesiones de entrenamiento específicas. Los resultados mostraron una gran cantidad

de ejercicios por tiempo efectivo que llevaron a una mejora significativa en la pronunciación en los estudiantes que entrenaron la herramienta CAPT desarrollada en cada prototipo. Además la retroalimentación ofrecida en el entrenamiento resultó ser efectiva, ya que aquellos que la siguieron obtuvieron mejores resultados que los que no. Finalmente, se encontraron altas correlaciones entre las puntuaciones subjetivas de evaluadores humanos y las objetivas del sistema ASR, tanto con la puntuación global de uso de la herramienta como la puntuación del post-test.

Estos prototipos experimentales se centraron en responder a las preguntas de investigación RQ2 (y sus *Issues*) y RQ1 (e *Issue 1.1*), tratando aspectos relativos a los objetivos RO1, RO2, RO3 y RO4. Los principales resultados del prototipo *English Vowels* han sido publicados en [23], [24]. Los principales resultados del prototipo *Japañol* han sido publicados en [25], [26], [27], [28], [29]. Un avance de la metodología que se llevará a cabo en el útimo prototipo, Estoñol ha sido publicado en [30]. En la Sección 7.4 se encuentra la descripción completa de dichos prototipos.

R1.6.4 Prototipo *COP*

A la luz de los resultados del segundo experimento, y casi en paralelo con el tercero, se llevó a cabo un nuevo experimento (***Competitive Learning***) con un enfoque de juego más competitivo, llamado (***COP***), para intentar evitar el estancamiento en la participación y en la mejora de la pronunciación detectado en el segundo experimento, al mismo tiempo que se obtenía una mayor cantidad de datos de habla y de comportamiento del usuario con dicho sistema CAPT. En particular, se desarrolló una nueva versión de la aplicación desarrollada para el segundo experimento (*TipTopTalk!*), en la que los alumnos podían desafiarse entre sí bajo un conjunto de reglas comunes —a diferencia de *TipTopTalk!*, en el que los usuarios jugaban solos). Los resultados mostraron una práctica intensiva respaldada por una cantidad significativa de actividades y días de juego. En concreto, los jugadores más activos y motivados en la competición lograron resultados de mejora de pronunciación significativos. Estos resultados quedaron respaldados con las respuestas a los cuestionarios y grupos focales llevados a cabo.

Por último, este experimento como el segundo, intentó responder a las preguntas de investigación RQ2 (e *Issue 2.1*, *Issue 2.2*, *Issue 2.3*) y RQ3, junto a la RQ1 (e *Issue 1.1*), tratando aspectos relativos a los objetivos RO1, RO2, RO3 y RO4. Los principales resultados han sido publicados en [31]. En la Sección 7.5 se encuentra la descripción completa de dicho experimento.

R1.7 Resumen de Conclusiones y Trabajo Futuro

El trabajo presentado en esta tesis doctoral nos permite responder a las preguntas de investigación propuestas con una respuesta afirmativa en relación a la aportación de evidencias sobre el uso de tecnología del habla, metodología de entrenamiento específica y elementos de juego motivacionales en herramientas móviles CAPT.

- Gracias al avance de la calidad de las tecnologías del habla, hemos sido capaces de incluir tecnología de reconocimiento y síntesis de voz de una manera no obstrusiva en las herramientas CAPT desarrolladas en la experimentación. De esta forma, dichas herramientas han servido como un instrumento útil, didáctico y complementario en SLA para la mejora de la pronunciación a nivel segmental.

- Las decisiones metodológicas llevadas a cabo en las diferentes versiones de las herramientas CAPT diseñadas y validadas en este trabajo han permitido medir la mejora de la pronunciación relativa de las personas que entrenaron con ellas.
  - Se han utilizado listas de pares mínimos elaboradas mediante un novedoso protocolo semi-automático propuesto en esta tesis doctoral, que tiene en cuenta la L1 y L2 del participante y la tecnologías ASR y TTS.
  - Se han incluido dichas listas en ejercicios de exposición, percepción y producción en diferentes modos de entrenamiento, sonidos e idiomas, en los que se ha utilizado tecnología ASR y TTS.
  - Se han empleado diferentes técnicas de retroalimentación correctiva que han demostrado ser útiles y efectivas. Con ellas, los usuarios han podido superar los ejercicios de entrenamiento propuestos; y nosotros hemos sido capaces de averiguar sus mayores dificultades en cuanto a sonidos y actividades de entrenamiento.
  - Se han reportado no solo resultados positivos de mejora de habilidades de percepción y producción de manera objetiva y subjetiva, sino que los participantes de grupos que utilizaron las herramientas CAPT lograron una mejora mayor que la lograda en los grupos de instrucción con el profesor en el aula.
- Por último, los elementos de juego han tenido una influencia positiva en la motivación, el rendimiento y el aprendizaje de los participantes en los diferentes sistemas CAPT desarrollados en esta tesis. En concreto, la competición de COP ha demostrado ser un factor motivacional positivo, especialmente para los usuarios más activos, cuya participación intensiva en el juego les permitió lograr una mejora significativa de la pronunciación L2 al final del experimento.

Además de las colaboraciones que se mantienen relacionadas con esta tesis doctoral, la gran cantidad de datos recopilados y que los resultados de esta tesis son satisfactorios, hay algunos aspectos que pueden mejorarse y dar paso a nuevas líneas de trabajo futuro, como son:

- Analizar y diseñar algoritmos específicos de reconocimiento de voz para la identificación de errores de pronunciación que permitan caracterizar el nivel de habilidad de pronunciación. Con ello, se determinará el conjunto de características clave obtenidas al correlacionar los errores de pronunciación con las valoraciones de expertos humanos para que un sistema de clasificación automática permita formular recomendaciones personalizadas sobre el modo y lugar de articulación de la pronunciación.
- Encontrar nuevas técnicas para adaptar el sistema CAPT al usuario de una manera más personalizada e individualizada, ayudará aún más a mejorar no solo sus resultados de mejora de la pronunciación, sino también su grado de motivación durante el entrenamiento de pronunciación.
- Analizar la relación entre el diseño del sistema CAPT, la estrategia seguida por los participantes durante el entrenamiento y sus resultados finales será útil para clasificar y predecir el comportamiento de los usuarios con dicho sistema.

Parte I

Introduction

Chapter 1

Introduction

Acquiring a proper communication level in any foreign language is mainly affected by intelligibility, nativelikeness, comprehensibility, and fluency of speech. However, traditional language learning courses and systems often focus on other language skill areas, such as grammar or writing. On the one hand, recent advances in speech technology have reported new approaches that improve significantly the performance of voice recognition and speech synthesis. Consequently, these technologies are integrated into state-of-the-art pedagogical systems for pronunciation training as complementary tools through applications for smart devices, allowing learners to use them continuously and autonomously. On the other hand, learning games have a remarkable potential for education, and in particular, for language learning. They provide an emergent form of social participation that deserves the assessment of their usefulness and efficiency in the learning process. Nevertheless, there is still scarce empirical evidence about the effectiveness of CAPT systems with speech technology.

This chapter discusses the feasibility of the topics mentioned above for pronunciation training, introducing the reader in the field, and showing both the context and motivation of this thesis work. Furthermore, the specific problems and challenges that lead to the objectives and research questions defined for this dissertation are identified.

1.1 Motivation

There is currently a growing demand on second language acquisition (SLA). A recent study at the end of 2016 informed that there were approximately 912 million second language (L2) learners worldwide [1], a seventh part of the global population in that year. Besides, communication between people from different places of the world is no longer a problem since the advancements in technology ease this process. However, the variety of languages and cultural environments of individuals might be a barrier to develop a successful communication.

The predictions for the year 2020 advanced that each person would have access to 6.58 smart devices [1]. These devices are present in e-learning and one-to-one tutoring, interesting alternatives to traditional in-classroom courses. In particular, computer-assisted language learning (CALL) and mobile-assisted language learning (MALL) systems integrate advanced technology that become very attractive to language learning and can help in the process of learning and teaching in an efficient way. Even though such devices and technology cannot serve as human tutors, they can perform a complementary role in education by increasing efficiency and

motivation of the learning process, being used anywhere at any time, and repeated as many times as desired.

Including game elements in educational tools favors a better individual performance [2]. Currently, the number of learning games for language learning is increasing. Recent studies report that learner's motivation and engagement are enhanced not only inside but also outside the classroom [3]. Although the inclusion of social and competitive elements in any pedagogical system must be done with caution, there are studies that indicate that competitiveness in the context of game-based learning facilitates the achievement of learning objectives [4] and encourages cooperation as an articulating element of class work [5].

Computer-assisted (aided) pronunciation training (CAPT) systems support innovative research and practices which lead to transform language learning, creating opportunities to revisit old ideas and challenge established beliefs [6]. CAPT is an important sub-area of CALL and MALL constantly undergoing change, which combines corrective feedback and automatic pronunciation quality assessment, among other functionalities provided by the speech technologies incorporated. They are often automatic speech recognition (ASR) and text-to-speech (TTS), which transform speech into written text, and vice versa, respectively. Nowadays, these systems are supported by enormous quantity of data and complex algorithms that improve their quality significantly. For instance, the Google company reported that their recent advances in machine learning applied to TTS have helped to generate speech waveforms 1000 times faster than before (generating one second of audio only takes 50 milliseconds), and have achieved over 20% better quality ratings than standard voices [7]. Besides, in the field of speech recognition, Google have also achieved a word accuracy rate of 95% for the English language, therefore reaching the threshold of human accuracy [8]. Better rates are likely to be obtained in the near future with the use of more sophisticated deep neural network (DNN) techniques and greater central processing units (CPUs) [9], [10]. Although until 2014 there were scarce peer-reviewed research investigations about CAPT (only a 26.9% of the 75 state-of-the-art studies surveyed in [11]), CAPT systems are being incorporated in recent experiments more frequently due to the improvements and new possibilities they offer [12].

A correct training methodology must adequately address the aspects of instant automatic feedback and the design of activities and teaching elements according to learner's L1 and L2, in order to optimize CAPT tools' efficiency and use time [13]. The L2 acquisition learning process is intensely affected by a well-established habitual perception of articulatory motions and sounds in the learner's mother language (L1). It often leads to mistakes and inaccuracy in speech production of the L2 learners (i.e., a negative language transfer [14]). For these reasons, in this thesis the minimal pairs technique has been used. Minimal pairs (pairs of words that vary by only a single sound) bear great benefits in pronunciation learning and teaching since they appear in almost all languages and can contrast L1 and L2 sounds [15]. They are often used for teaching L2 segmental pronunciation (i.e., the teaching of single speech sounds, such as vowels, disregarding intonation, and other suprasegmental aspects of connected speech [32]).

In summary, L2 learning, and more precisely, L2 pronunciation training, is opened to new teaching paradigms. The possibility of learners training anytime anywhere,

at their own pace, allows teachers to provide small group and individualized instruction rather than lecturing to an entire class. The fact that today's students are digitally literate motivates the development of learning CAPT systems for smart devices. Learners are used to interact into digital environments for communication, information, social contact, gathering, and analysis. Although they might be digital natives, comfortable with, and immersed in technology, they depend on teachers and experts to learn through digital means. Besides, ASR and TTS technologies have drastically enhanced their performance in recent years, being possible to be integrated into educational resources. Therefore, the current challenge is to carefully design and adapt an effective CAPT system with not only such technology, but also with a training methodology, a pronunciation improvement assessment, and a corrective feedback strategy, according to learner's L1 and L2.

1.2 The Problem

Providing an effective set of pronunciation training activities is not an easy task since there are several factors to take into account:

- **Learner's L1**. The similarities and differences between mother and target languages vary the difficulty of the learning process.
- **The set of personalized training activities**. Depending on the student's L2 pronunciation level and her/his performance, the activities recommended must be individualized and adapted to each learner in order to be effective.
- **Assessment of learner's results**. Both subjective and objective scores can be provided to users, not only at the end of the experimentation, but also during the training.
- **The feedback provided to the students**. More feedback than a teacher alone can give in class is needed. However, in some cases this feedback is insufficient or too difficult to understand by the learners.
- **The technology included in the training methodology** must be carefully selected since these assessment scores must be as precise as possible.
- **Motivational elements**, such as game elements or interaction with other learners might influence the results of the training, deviating students from the actual pronunciation improvement goal or even discouraging them.

A research challenge that raises the problems described above must be addressed from a multidisciplinary perspective which includes: learning methodologies, design of learning tools, data-based modeling, and evaluation techniques, among others.

Finally, establishing proper protocols for gathering and analyzing results from the experiments is necessary since the effectiveness of the CAPT tool is influenced by its scalability and performance. Firstly, the possibility to extend the range of languages leads to generalize training concepts and strategies, and to correctly select the necessary speech technology. Secondly, the user's interaction with a CAPT system tends to be massive in terms of speech data and log activity. In some cases an important investment of money is necessary in devices and computer servers.

1.3 Objectives and Research Questions

The main objective of this thesis is defined as:

To design and evaluate a CAPT tool for smart devices which incorporates **current TTS and ASR technology**; helping students to **work autonomously**, at their **own pace**, and with the possibility of providing **real-time feedback**.

This main objective 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 TTS and ASR technology.
- **RO2**. To select the most appropriate metrics for the assessment of the speaker's pronunciation level.
- **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 speaker's L1 and to the limitations of the TTS and ASR technology.
- **RO4**. To select and design a CAPT system with current TTS and ASR technology that provides an individualized feedback to the speaker for improving L2 pronunciation.

In order to carry out the experimental procedure of this thesis, **three research questions** are identified to validate the research objectives, categorized by topics. The first topic is related to the *feasibility of current speech technology (TTS and ASR systems) integration in CAPT tools*:

- **RQ1**. Can current TTS and ASR systems be successfully used in a non-obstructive way in the CAPT tool developed?
  - ***Issue 1.1***. Can current TTS and ASR systems help to assess different groups of speakers according to their L2 pronunciation level in the CAPT tool developed?

The second topic refers 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 student'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 and most difficult training activities)?

Finally, the last research question aims 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?

1.4 Research Methodology

In order to accomplish the objectives and give answers to the research questions proposed in this thesis, an **experimental research** [16] is conducted for the whole experimentation process with a multidisciplinary group of researchers and experts. Five phases can be defined in each experimental iteration:

1. **Identifying the research problem**. The process starts by clearly identifying the problems that will be addressed during the research process, starting with the existing solutions in the state-of-the-art, and considering what possible methods will affect a solution.
2. **Planning the experimental research study**. An experiment is carefully devised to test the research objectives and questions.
   (a) **Selection of participants**. The target population, enrollment rules, sample size, and groups are defined.
   (b) **Variables**. Different metrics are defined to measure the research variables from the data results gathered from the instruments.
   (c) **Assessment protocol**. The research variables are measured before, during, and after performing the training activities.
   (d) **CAPT tool development**. For each experiment in this thesis a novel CAPT tool is developed.
3. **Conducting the experiment**. At the beginning, the participants' groups must be established. Then, each user performs the activities defined for her/his group in the previous phase, and the experimental data related to the variables of the study is collected with specific instruments for each experiment.
4. **Analyzing the data**. The data gathered is analyzed. It must be decided which indicators will be, and will not be, important, in order to corroborate how the experiment is successful.
5. **Publication of findings**. The most relevant results are shared and published in scientific journals and conferences by means of articles, abstracts, show and tell demonstrations, and presentations.

1.5 Outline

This document is structured in six parts. In the first chapter of the first part (Chapter 1), the main topic of this thesis has been presented and motivated, the research objectives and questions have been settled, and a global vision of the research methodology carried out for this dissertation has also been given.

In the second part, a deep revision of related work in the state-of-the art is presented and discussed on the light of the main characteristics of systems and experiments related to the objectives of this thesis. In particular, in Chapter 2 a review of traditional pronunciation training activities and CAPT with the possible integration of ASR and TTS systems in pronunciation instruction are reviewed. In Chapter 3 the state-of-the-art pronunciation improving assessment strategies for CAPT are discussed. In Chapter 4 the corrective feedback strategies adopted by the state-of-the-art CAPT studies are examined. In Chapter 5 the fundamentals of individualistic

and social learning applied to pronunciation training are described. The implications of gamification elements in learning contexts are also stated.

The specific details about the experimental framework's concepts, strategies, and elements necessary for the experimentation are specified in the third part of this document. In particular, in Chapter 6 an exhaustive review of the common dimensions of the experimental procedure is included, in relation to the state-of-the-art. In Chapter 7 each experiment of this dissertation is detailed in depth, giving an evolutive vision of the work carried out along this thesis. In Chapter 8 the results obtained in all experiments are discussed to give answer to the research questions of this thesis.

In the fourth part (Chapter 9) the conclusions are summarized and some future directions of this thesis work are defined. Furthermore, the publications, research funding, achievements, and attributions obtained during the course of this dissertation are enumerated.

The appendices are included in the fifth part of this document. Appendix A explains in depth the algorithm for elaborating minimal pairs lists designed in this thesis. In Appendix B an overview about the similarities and differences of all the experiments of this thesis is presented via comparative tables. Appendix C shows the pre-test and post-test list of words given to users in the experiments which included them. Appendix D represents the main characteristics of the speech corpus data gathered in two experiments of this thesis. Appendix E sheds light to a standard Kaldi project structure and a list of steps for developing an ASR system. Finally, in the sixth and last part of this thesis the references in the bibliography chapter are included. They are compliant with the IEEE Reference Guide[1].

[1] http://ieeeauthorcenter.ieee.org/wp-content/uploads/IEEE-Reference-Guide.pdf

Part II

Literature Review

Chapter 2

CAPT Methodologies for Second Language Learning

Speech is the first and universal means for information transmission among human beings. Since our natural way of expression is oral language, we are ready to acquire and use it. In fact, an intelligible pronunciation when talking with someone in a different language is key to achieve a suitable communication level in our job, our trips, or in our daily life. Therefore, recent advances in technology rise the interest in adapting computers and mobile (smart) devices to use speech as a way to present and receive information.

Teaching styles have changed remarkably over the years and in order to reach as many students as possible, teachers need to create a diversity of learning experiences. Integrating technology into the classroom is a simple way to meet the needs of varied learners, allowing students to learn by doing and at their own pace.

Speech technologies have undergone a great development in recent years. The high similarity indices of synthetic voice with natural-sounding speech, and the almost human reliability of speech recognition allow the possibility of their integration in CALL and MALL systems for the teaching of L2 pronunciation. In this particular case, they are defined as CAPT systems. Besides, technological advances of computers and smart devices over the last years have allowed technology not only to be used in classroom, but also at home. Although a CAPT system is not meant to replace a teacher, it accomplishes similar functions to those of a competent teacher and recommended teaching material —to a certain extent.

To date, there have been very few attempts to empirically measure CAPT effectiveness in pronunciation training courses and experiments, much less in applications for smart devices. In fact, traditional language learning (LL) courses and techniques are mainly focused on other linguistic competences, such as grammar, semantics, or vocabulary. Interaction with a teacher or other classmates with written and audio materials are the typical activities in traditional pronunciation training learning courses. However, a high number of students or activities can be turned into a bottleneck in time and resources. This type of activities can be performed in CAPT systems with speech technology. The success is directly related to a correct choice of training methodology (set of activities and feedback). It is necessary to adapt the methodology's elements correctly to reach a proper set. CAPT systems allow students to perform a high number of activities at their own pace, in a stress-free environment, and to receive individualized and instantaneous feedback. Such intensity of guided individual work is hardly ever attained in the average classroom.

The training methodology of this thesis is focused on training and improving segmental pronunciation with a specific set of methodological choices, such as L1–L2 connection, particular lists of minimal pairs, an exposure–perception–production training cycle, and the inclusion of ASR and TTS technologies.

In this chapter, the main training activities performed in pronunciation instruction and their implication on SLA are pointed out. Then, the reasons to use speech technology (ASR and TTS systems) in L2 pronunciation teaching with CAPT are argued by providing (1) an overview about the main features of CAPT systems; (2) a detailed revision of the literature about speech recognition for CAPT; and finally, (3) an explanation about the evolution and main experiments of CAPT with speech synthesis.

2.1 Pronunciation Teaching in Second Language Acquisition

The research field of SLA is experiencing significant changes as a consequence of the recent interest in L2 pronunciation in the last two decades [11], [33]. New methods, questions, and scholars are appearing since the use of technologies (CALL) has accelerated these changes [34], [35]. However, the specific field of pronunciation has occasionally been faced to historical forgetfulness, such as old ideas as new revelations and claims that haven been clearly refuted in the past, such as the very concept of the intelligibility principle in 1900 [33]. Giving more details about the historical evolution of pronunciation training exceeds the limits of this work. Interested readers can find an interesting historical review in [33].

Training L2 pronunciation permits speech intelligibility and comprehensibility to be enhanced, fluency to be improved, and it is a means to achieve a native-like proficiency in all aspects of L2 [11], [33]:

1. **Accentedness (nativelikeness)**: perceived differences in pronunciation as compared with a local variety, focusing equally on all pronunciation features in an L2.
2. **Comprehensibility**: how easy L2 speech is for a listener to understand.
3. **Intelligibility**: how understandable L2 speech is.
4. **Fluency**: fluidity of speech (absence of dysfluencies, such as filled and unfilled pauses, self-repetitions, or false starts).

Although these four focuses of pronunciation are related, they are usually trained and analyzed individually by means of a specific scope of training —segmental, suprasegmental, or both—. The recent studies of this field in the literature currently target intelligibility over the traditional nativelikeness [11], [33].

Regardless the focus, pronunciation training in classical classroom instruction consists of two main phases. First, the interaction with a teacher/instructor, which in most of the cases means to listen to a model speaker, interspersed with learner's repetitions aloud. Second, and this can be seen as an addition to teacher-interaction, tasks that are performed either at home or school by the individual learner. These exercises are usually based on written material (i.e., books or exercise sheets) and sometimes in recorded audio material. Monologue and interactive situations with interactional partners are also possible.

Listening and speaking skills are taught and trained by means of perception (hear new sound contrasts) and production (utter sounds). The following training tasks are mainly used to assess perception in both segmental and suprasegmental level in pronunciation teaching [33], [36]:

1. **Identification**. Hear a segment, word, or other unit, and select the written unit or image to which it corresponds.
2. **Discrimination**. The learner must decide if two units are the same after hearing them. Other variant consists in hearing two units and decide if a third unit corresponds to the first or the second one.
3. **Oddity**. The learner must identify a syllable, word, or longer utterance that differs from at least two other. If there is no odd word, the learner must indicate that all words are the same.
4. **Matching**. Subjects must hold the pronunciation of a unit in memory while they determine whether the pronunciation of the next units includes/is the same or not.

On the other hand, the following training activities are typically used to assess production [33], [36]:

1. **Oral reading**. See a word or sentence and read it aloud.
2. **Oral repetition**. Hear a word or sentence and repeat it.
3. **Picture naming/description**. See a picture and name/describe it in a few sentences.
4. **Picture narration**. See a series of images that tell a story and narrate it.
5. **Monologues/dialogues/interactions**. A long speech by one or more learners.

The main findings related to the link perception-production in pronunciation training in SLA are [33], [35], [36], [37]:

- Subjects find more difficulties in production tasks than in perception ones.
- Accurate perception is a necessary, but in some cases insufficient, condition for accurate production.
- In specific contexts of learning, such as an instructed or classroom context, both perception and production will probably change at different rates, with production improving after perception.
- Sleep and rest seem to enhance perceptual learning, in particular when training occurs immediately beforehand.
- Individual differences of subjects associated with more accurate perception and production may also influence the perception–production connection.
- The absolute improvement of lower proficiency subjects is higher than the improvement of higher proficiency subjects after training.

Although the specific implementation and examples of both perception and production activities in the experiments of the state-of-the-art exceeds this thesis work, only those experiments which include CAPT are detailed in Section 2.2 and its subsections.

2.1.1 Pronunciation Teaching at the Segmental Level

The extent to which learners can direct their attention to phonetic forms is also important [37]. Segmentals have been studied more often than suprasegmentals in the literature, being English and Spanish the main analyzed languages as L2 [11]. Most of the studies reported moderate positive correlations between L2 segmental perception and production for speakers of an L2 [38]. Furthermore, the relationship between foreign accent and age is undeniable, being also clear the L2 prediction strength of accent among learners of a similar age can be predicted by their degree of exposure/experience with the L2.

The magnitude of differences between L2 learners and monolingual native speakers of the target L2 depends, partly, on the degree of perceived phonetic dissimilarity of L2 sounds from the closest L1 sound [38]. For instance, it might be easier to learn the L2 phonemes /x/ and /y/ if they resemble the [x] and [y] allophones of two different L1 phonemes than if the L2 /x/ and /y/ phonemes resemble the primary allophones of an L1 phoneme [38].

A novel and innovative method for segmental pronunciation training called **Native Cardinality Method**, **NCM** [39], [40] takes the native L1 phonological system of the student (Spanish in this case) as a starting point, and follows an intensive cyclic training protocol of L2 training that includes three phases:

- Articulatory knowledge (**exposure**).
- Perceptive awareness (**perception**).
- Sound realization (**production**).

The NCM follows some of the activities with minimal pairs presented in other related training programs [41], [42], [43]. In particular, it uses mixed (L1–L2) minimal pairs and approximate pairs, introduced at different stages of the teaching–learning process, and tries to avoid or, at least, reduce the transfer of Spanish pronunciation to English language. A mixed minimal pair consists of a Spanish and other English word which differ in only one sound (i.e., *su* (Spanish) – *Sue* (English)). The latter group is formed by pairs that differ in more than one sound (i.e., the first two sounds of *ten* (Spanish) – *den* (English) are different). Even when there is no possibility of finding minimum mixed pairs in this particular context, the author of this method uses quasi-neologisms, artificially created words that, having a graphemic and phonemic dimension, lack a semantic dimension. That is, they are invented words which do not mean or claim to mean anything, but they are useful to compare Spanish and English sounds. For instance, *sam*, does not exist in Spanish, but it would be a diminutive of *Samuel* – *Sam* (English).

2.2 Computer-Assisted Pronunciation Training

The discussion about the Information and Communications Technology (ICT) contribution to language learning versus traditional teaching methods has increased over the last decade since research has investigated the ways in which technology can improve pronunciation training [11], [35]. CAPT is comparatively much more recent than CALL since the first research studies appeared in the late 1990s [11]. Furthermore, CAPT systems have evolved from desktop tools to applications for smart devices [44], [45].

Following the description of a CAPT system in [46], it can "provide learners individualized instruction, frequent practice through listening discrimination and focused repetition exercises, and automatic visual support that demonstrates to learners how closely their own pronunciation approximates model utterances". That is, learners can access automatically and at their own pace to an unlimited pronunciation training practice (the same tasks described in Section 2.1, but with technological help) with automatic measurements.

Nowadays, CALL and CAPT are able to provide many benefits to teachers and learners, including stress-free and interaction-rich context where teachers enjoy more opportunities to address individual needs of students, since not all situations can be previsioned and programmed in a computer application; while the students can practice at their own pace and get immediate personalized feedback [34].

Running an L2 CAPT project involves a long-term collaboration among a varied range of specialists from diverse areas (i.e., phonetics, computer science, linguistics, pedagogy, and engineering). Establishing a common understanding of the multidisciplinary project goal in order to be successful is desirable.

CALL and particularly CAPT can be considered not only as an optimal complement to a teacher-based class but also as the core technology of an entire language learning course. Although existing literature on CAPT indicates that these systems tell us how good learners perform and how to improve their pronunciation [47], [48], choosing the correct methodology and elaborating a sound experimental validation of this technology's pedagogical effectiveness assessment, are necessary to further develop the field [13], [49]. Speech technology can help individuals to improve speech perception and speech production skills by raising awareness for phonological contrasts through exercises of discrimination tasks with speech synthesis technology and those involving the user's own speech with the help of speech recognition. Thus, there is an immense potential for speech technologies to be used in language learning scenarios since they provide quality and availability.

CAPT systems may offer learners a non-obstructive and stress-free environment, in which students can practice at their own pace and access practically unlimited input [50]. While some CAPT investigations apply technology in innovative ways (i.e., including ASR or TTS systems), others emulate traditional (i.e., non-computerized) classroom procedures [11]. In addition, the majority of current CAPT systems propose isolated exercises, such as perception, production, or exposure ones, within learning courses.

2.3 Speech Recognition in Language Learning

Due to the advances in signal processing, algorithms, architectures, and computing platforms, speech recognition has undergone a great development. The first speech recognition device, Audrey system, built by Bell Labs in 1952, recognized only ten digits spoken by the same single voice. Nevertheless, it is not until the 1990s when speech recognition makes sense in LL. Different dictation software programs started to market, such as Dragon Naturally Speaking, IBM Personal Dictation System, and Kurzweil Voice. Furthermore, the demand for speaking practice first, and in speech therapy later, increased notably.

There was also a growing wave of interest in the one–to–one tutoring [51]. However, a human tutor for every student was not feasible. It entailed the appearance

of the first computer tutors and the employment of ASR for LL and speech therapy. Although Dragon Naturally Speaking was designed by native speakers and was not developed for error detection, [52] reported that it might have pedagogical value in the future as a means of giving corrective feedback and identify the problems in pronunciation that affect humans' understanding of non-native speech. [53] complemented the idea of [52], by affirming that ASR's accuracy must be tested by natives (in this case English). In particular, he proposed that when a more highly developed version of the software incorrectly recognized a word, students might view the computer's error as an indication of a mispronunciation which needs correction.

Since late 2000s, there has been a growing interest in CAPT by means of ASR because these systems provide learners automatic and individualized feedback in a private environment [13]. Two main tendencies in ASR for LL can be distinguished. First, computer-desktop applications for pronunciation training [11]. Second, the possibility of learning anytime and anywhere with a diverse range of smartphone applications makes possible learning not only individually but also collaboratively and competitively (i.e., Duolingo[1], Babbel[2], or ElsaSpeak[3]). This type of applications often turns into online courses [54], [55].

The core ASR systems integrated into the mentioned applications varies from open source frameworks with a great community support, such as CMU Sphinx, Kaldi, and Julius to commercial and very powerful ASR systems, such as HTK, Dragon Dictation, Google Now, Cortana, Siri, or Alexa, among others. Interested readers can find a detailed description about speech recognition fundamentals, characteristics and examples of ASR systems in Section 6.4.1.

Choosing a proper ASR system is not trivial. The objective of the study and the subjects target must be taken into account:

1. **ASR systems for native speakers**. There are some characteristics to take into account, such as pronunciation variation, end-point detection, disfluencies, and background sounds, among others. Examples of this type of ASR systems are dictation software applications, personal assistants, or voice command systems.

2. **ASR systems for non-native speakers**. Systems more complex than the previous group which often have a degraded performance. They include atypical and pathological speech (i.e., a corpus populated with utterances of non-native speakers). The three main knowledge sources of ASR systems are clearly affected (grammar, set of words, and pronunciation deviations). Besides, there are problems in read speech (bad linking of words) and in spontaneous speech (more filled pauses).

There are several inclusive strategies to improve non-native ASR systems' performance. The first approach and more generalist one, is to optimize the acoustic models, lexicon, and language model to compensate for deviations. Another option is to restrict the search space. For instance, constraining learner's output by elicitation strategies, such as reading aloud and repeating auditorily prompted sentences.

[1] http://duolingo.com
[2] http://babbel.com
[3] https://elsaspeak.com/en/

On the other hand, ASR systems performance can be improved by altering other external factors. First, analyzing what has been said, adjusting the grade of tolerance. Second, examining how has it been said (i.e., error detection and find deviations from typical speech). Finally, providing feedback to the learner. In summary, an ASR system must be optimized not only internally (system design) but also externally (technology, context, and train/test data).

2.3.1 ASR-based CAPT Systems in LL Experiments

Empirical experiments in LL are scarce, and almost non-existent for CAPT systems in mobile devices even though ASR-based pronunciation training has several advantages, such as dynamic evaluation, individualized feedback, more intensive practice, anxiety-free context, and opportunities for repair [50], [56].

However, erroneous feedback could be given: false positives and false alarms [57]. Four categories of ASR errors that could be used as predictors of L2 learners' difficulties are identified in [58]: homophones, minimal pairs, breached boundaries (in the context of two or more linked words), and negative cases. In phonology, a pair of words is said to be minimal (minimal pairs) when they differ in only one segment (sound) [59]. Learners carry out virtual risks of producing wrong word meanings when the correct phonemes are not properly uttered by simply altering a single segmental element. By simply altering a single phoneme, learners risk producing unwanted meanings. The distinction between both words is, *a priori*, a tough task for ASR systems due to the phonetic distance between words being very small. Finally, negative cases, such as can/can't or legal/illegal are also important for ASR since their misrecognition can thoroughly change the meaning.

There are some CAPT experiments which include theoretical lessons and activities related to them. In the experiment reported by [13], four training lessons were presented to the participants. In each of them an explanatory video was shown first, and then 25 sequential and guided exercises based on the video were proposed. Those exercises could be written questions to be answered orally by recording one of several possible answers or requiring the student to pronounce specific words for which example pronunciations are given. A similar training method is applied in [60]. First, a teacher explained the position of the tongue for the phonemes /r/ and /l/. Then, the training of prolonged /r/ and /l/ was administered by using spectrographic representations with overlaid formant-tracking results. The ASR system showed, in real time, the hidden Markov models (HMM) scores obtained in the production of the minimal pairs. The experiment described in [61] used 23 exercises of increasing difficulty. Each one of them emphasized the contexts: placement test, vowel/consonant, only one word or a sentence, and anticipation. It is important to note that this software offers six different hints in order to improve incorrect pronunciations: oneself and native exposure, instructions of how to pronounce the sound, image of side headcut with the position of the lips and listening to the word in a minimal pair or sentence. In [62], an HMM ASR-based CAPT system called *PLASER* presents 20 lessons, teaching two phonemes in each one. The exercises in each of the lessons are: (1) read-along: no assessment; (2) minimal pair listening: ear training; (3) minimal pair speaking: produce one of the words of the pair; and (4) word list speaking: produce one of the words of the list.

In [63], [64], the *Talk to me* software provides six dialogue sequences (each one has thirty question–and–answer screens), where the program asks a question to which the user responds by uttering one of three answers presented on the screen. There is also optional explanatory feedback of sound articulations. The difficulty level of the

speech recognition can be adjusted to require a looser or tighter match to the underlying models. In [65], the *PARLING* system is reported. This is an HMM-CAPT system that sets learners the task of making up a word-based story through word game activities and the possibility of creating their own dictionary. The methodology of word production is simple: the user's utterance of the word is accepted/rejected by the system as the answer. It also offers the possibility of listening to a native recording of the word. In [66], the *Nuance Dragon Dictation* software, a speaker-independent dictation system designed for continuous speech recognition installed on students' mobile devices, gives immediate feedback to students when they read aloud the target words and phrases in French in 20–minute pronunciation activities. In [67], students interact with the *Moby.Read* application. In each test session learners read (1) a word list, (2) an easy practice passage, and finally (3) three grade–level passages. After that, they are asked to retell the passage in their own words, with all the details possible and then answer two short questions aloud. Scores are provided in real-time.

Finally, there are other studies that promote training methodologies over the Internet, such as peer-reviewing of read sentences after tasks of reading and speaking with peers [68], or conversations with natives or other L2 learners after the training with minimal pairs and lists of words in activities of perception and native imitation—production [69].

2.4 Text-To-Speech in Language Learning

In the early 1980s, Texas Instruments' Speak & Spell built the first single-chip voice synthesizer for a toy –called Spelling Bee– to teach children how to spell. The literature about the TTS benefits for pedagogical applications is very limited and almost non-existent for mobile devices contexts. It was only recently that some studies confirm TTS advances seem to be ready for use in LL activities [70], [71], [72].

The use of TTS systems as part of pedagogical tools has generated a great controversy and, there is still certain debate about their suitability for L2 pronunciation training (like ASR systems) and very few attempts to empirically measure their performance. However, recent research in speech synthesis has reported some benefits in terms of comprehensibility, naturalness, accuracy, and intelligibility [70], [72], [73]. TTS systems can raise learners' awareness of certain language features in a personalized and learner-centered way [74]. In particular, their sound quality is adequate to be used in the generation of pronunciation models of phonemes, words or sentences for helping students to improve their discrimination and production skills.

TTS systems are suitable in terms of promoting some of the ideal SLA conditions presented by [75], [76], such as learner fit, authenticity, potential for providing feedback, and learning strategy development. Interested readers can find an overview about speech synthesis and the features taken into account in this thesis to choose a TTS system in Section 6.4.2.

2.4.1 TTS-based CAPT Systems in LL Experiments

The importance of listening to sounds before producing them is contrasted in [77], who said that the learner's brain converts an unclear sound into the closest sound found in L1, and suggested that emphasis should be put on listening. This importance has also been accented by [37], [78], [79].

There are scarce experiments in LL which integrate speech synthesis due to controversy produced by the limitations of TTS systems [73]. However, the quality of

these systems has currently increased due to an even larger corpora and new statistical parametric and DNN to process both superficial [80] and HMMs [81] information from the mentioned corpora.

Traditionally, the method called High Variability Phonetic Training (HVPT) consisted on exposing learners to multiple natural voices producing the target sounds, rather than a single voice (i.e., teacher's voice in the classroom). Then, learners only had to choose the word that they have listened to (a single task). Studies with minimal pairs [82], [83], [84] and with single words [37], [85], [84] report learners' perception and production improvement. Nowadays, this method currently can integrate different synthetic voices.

There is even less research in speech synthesis for L2 CAPT systems, since most of them, to date, have used natural-speech as a model to listen to, imitate, and self–compare [86], [87], [88], [89], [90]. Imitation procedures have also been applied to pitch and intonation suprasegmental forms in sentence production [91], [92], [93], [94] or global speech characteristics [78], [95]. Some recent methods use manipulated natural-speech recordings in order to assist the identification and discrimination of individual phonemes, improving also production [83], [96].

Regarding the few experiments existing with speech synthesis in language learning, in [72], the majority of participants who listened to speech samples, alternately produced by TTS and a human, reported that TTS technology could and should be used as a tool for LL to perform perceptual activities with both natural and synthesis speech; whereas in [97], the *NaturalReader* TTS software system allowed students to complete weekly pronunciation tasks in a computer. They consisted of listen–and–rank, listen–and–categorize, and listen–and–repeat sentences. They were asked to fill in reports results manually after training.

2.5 Summary

CALL has significantly contributed to new changes in SLA. In particular, CAPT systems are intended to be a useful resource in pronunciation training due to the emergence of new technologies and services for smart devices, and the unceasing growth in demand of both, off-line and online L2 courses. However, there are not yet enough empirical experiments on mobile CAPT systems that resort to ASR and TTS technologies nor reports on their effectiveness.

In this thesis, off-the-shelf ASR and TTS systems are incorporated into different versions of mobile CAPT systems in a non-obstructive and user-friendly way, allowing designers and experts to personalize instructions for learners, and saving time and resource costs. They also offer the possibility of assessing different language level users.

In this chapter, a general overview of pronunciation teaching in current trends, training activities, and findings has been described. The reasons why ASR and TTS technologies can be integrated into personalized mobile CAPT systems have been also analyzed. Then, the set of methodological decisions of the most relevant CAPT tool of the state-of-the-art have been reviewed. In particular, an exhaustive revision of CAPT experiments of the literature has been detailed. Firstly, the evolution of speech recognition in LL has been described, pointing out the most relevant experiment with ASR-based CAPT systems. Finally, the main features of speech synthesis systems in LL have been described. The experiments in the literature about TTS in LL contexts have also been pointed out.

Chapter 3

Assessment of Pronunciation with CAPT Systems

Replicating a scientific experiment is crucial to ensure its validity. In particular, a clear statement of the assessment methods to process the obtained results leads to an increase in their significance and confidence level. Although the number of experiments with CAPT systems based on mobile speech technology is increasing, there is not yet a common protocol respecting the evaluation of the improvement in pronunciation when using them. In particular, there are very few contributions on assessment of CAPT's pedagogical effectiveness with speech technology, and in some instances it remains unclear.

On the one hand, the pre/post-test and pre/post-quest designs are the preferred subjective method to compare different groups of learners, gather their opinions, and measure the degree of change of their perception and production skills after specific training sessions. The main limitation of this approach is the possible inconsistency of scores provided by human raters, probably originated by the lack of common guidelines, and the fatigue that such an evaluation involves. On the other hand, several quantitative and very specific metrics obtained from ASR for pronunciation assessment are proposed in each one of the limited state-of-the-art experiments.

In this thesis a mixed assessment approach for pronunciation training in CAPT systems is proposed. In the first stages of the experimentation, experts provide subjective measures of learner's utterances. Then, these scores are correlated with objective and quantitative ones obtained automatically from the CAPT tool. The higher the correlation achieved, the greater the level of confidence scoring. The aim is to be able to automatically evaluate speakers in real time when practicing with the CAPT system. This assessment can also be applied to pre-test and post-test activities, in addition to a final CAPT system's score. Thus, future experiments can rely on automatic and objective scores provided by the technology that can serve as support when it is not possible the human help, saving time and resources.

In this chapter, the subjective techniques of pronunciation's improvement assessment after using a CAPT system are described in the first place and a revision of the literature about experiments which apply these methods is carried out. Then, a description of the metrics used to assess pronunciation quality objectively with examples of the state-of-the-art is presented. Finally, experiments that combine both approaches are reported.

3.1 Subjective Assessment

Human ratings offer a subjective approach for pronunciation activities. A 79% of the experiments surveyed in the effectiveness of L2 pronunciation instruction review in [11] based their results in this technique. A preparatory session among the human raters for sharing common aspects for evaluation is recommended since scores could be inconsistent and unclear. These individuals can be native or specialists in the L2–target language. In particular, the most usual protocol reported in the review consists in comparing the pre-test scores with the post-test ones in different groups of participants (i.e., control and experimental). In some cases, a delayed post-test was also carried out to ascertain the long-term retention. These tests generally contained the same stimuli (i.e., the same words to discriminate in perception exercises or to utter in production ones). Different rating scales were also used (i.e., right/wrong/no response, scales from 0 to 3/10/100, among others).

Some works report pronunciation improvement with perception and production activities of isolated phonemes (segmental level) with minimal pairs in a pre/post-test design [82], [83] or with isolated words [88]. As for the suprasegmental level, there are some studies that, instead of using minimal pairs, use lists of words read aloud to evaluate perception skills [85], production ones [37], [86], [90], [93], [94], or both of them [89], [96].

Furthermore, there are experiments that analyze either the perception or production of sentences with numerical scales. In particular, linguistic functions of prosody (elements of speech that are properties of syllables and larger units of speech), such as marking the location of pauses, the stressed words, and the direction for sentence-final intonation are some activities for these perception tasks [78]. Phrases productions are analyzed in [68], [93]. Specific parts of phrases are assessed in [87]. Prosody is also evaluated with spontaneous speech production tasks in [91], [92].

There are other different approaches, such as the assessment of audio recordings by other classmates [98], oral presentations by experts [95], the analysis of spontaneous conversations, also assessed by experts [69], [87], [89], and the perceptual evaluation of the regenerated audio signal before and after transformation [99].

On the one hand, there are scarce studies about pronunciation improvement measurement with ASR-based CAPT systems. The production improvement achieved in [60], [62] follows a pre-test, training sessions and post-test design with an ASR-based CAPT tool and minimal pairs. Word and sentence-level perception and production activities improvement is numerically evaluated by human raters in [66], [100]. Production improvement of spoken words is analyzed in [13], [61], [65], [97]. On the other hand, there are even less studies about TTS implication in pronunciation improvement in CAPT tools, since they are beginning to appear. In particular, a combination of natural and synthetic speech is used to listen to minimal pairs in the experiment previously mentioned in [83]. Fully synthetic sentences are presented in [97], and the pronunciation improvement is assessed with a pre-test, post-test, and delayed post-test design. Finally, websites with TTS technology promote the self-study and are reported to be useful to improve pronunciation in a pre/post-test evaluation with human raters [101].

3.2 Objective Assessment

Despite the accuracy and preciseness of human ratings, large CAPT experiments with a great number of participants become into a tough task of assessment in terms of time and resources. A powerful alternative is to automatically generate scores with technology. This second approach of assessment consists in providing objective measures and optionally, to correlate them with the scores of human raters [11]. On the one hand, the interaction with a CAPT system can be assessed in real-time during training. For instance, a numerical score can be provided after performing perception exercises; whereas in production tasks, a text with the recognized speech and a numerical score can be shown. On the other hand, pre-test, post-test, and delayed post-test tasks can be also assessed with technology, in a similar way as the previous case, but asynchronously.

Even though there is a limited literature on objective assessment of CAPT systems, it can be categorized into three categories. First, objective measures can refer to intelligibility (acceptable/unacceptable scores); second, to quality (Goodness of Pronunciation, GOP-based scores [102], [103]) and finally, to nativeness-like (i.e., pitch contours, accent ratings, ASR-based confidence scores, among others).

A better correlation between human ratings and pronunciation scores at a sentence(s) level based on both, L1 and L2 language characteristics of learners instead on only L2 ones, is reported in [104]. These automatic scores are obtained on Gaussian mixture models (GMMs) log-likelihood and HMMs confidence scores. A phone-level comparison with a likelihood-based GOP is carried out in [13], [102], [105]. The production mistakes are assessed by comparing native speech to non-native one.

Regarding studies with ASR-based CAPT systems, in some experiments an automatic right/wrong assessment is given after a user's utterance by highlighting the wrong part (*Dutch-CAPT* system) [13] or showing the speech recognized (*Nuance Dragon* system) [66] without presenting to the learner any score or quality value. Another investigations show objective scores from an HMM-based ASR software, *PhonePass* [63]. In a posterior study, these scores are correlated to human rater ones [64]. A large correlation between human's and ASR scores of orally produced words in sentences is also reported in [67]. Finally, diverse ASR system outputs are adopted for the assessment of basic English vocabulary in young children [106], [107]. Scores provided are based on phoneme-level language modeling and prove they can be used to obtain good classification results, even with a relatively small amount of acoustic training data.

3.3 Summary

Pronunciation assessment in CAPT systems does not follow a common pattern. Despite the scarce number of studies about this topic, there are several evaluation techniques depending on the availability of human raters, the technology employed, and the scope of training. In this chapter, a detailed revision of the literature about assessment of CAPT's pedagogical effectiveness has been conducted.

One of the main contributions of this thesis is to provide an automatic scoring method for CAPT systems with speech technology, based on user's results. This score can be obtained during and after training. It can also save time and resources to researchers and teachers when the number of learners is considerable.

Chapter 4

Corrective Feedback with CAPT Systems

Corrective feedback (CF) in CAPT refers to the answers provided by the system when the learners make linguistic errors in their L2 pronunciation. Giving a proper CF is key to a CAPT system to be successful in its role of helping students to improve their pronunciation and increasing the effectiveness of the learning process. It also permits a more intensive and individualized practice within an immediate and anxiety-free context.

CF can be explicit if the learner is informed of the corrected form of the error, or implicit, otherwise. Although there are several studies about CF in SLA, there is not a common framework with clear guidelines to follow in the field of CAPT. Results reported show different methods, variables and definitions that lead to mixed —and not always comparable— outcomes. Several techniques are carried out, from the easiest ones, such as reducing the result to correct or incorrect, to more complex methods, such as showing spectrograms and vocal tract videos. However, in most cases the feedback offered by these tools is insufficient or too difficult to understand by the users. Besides, teachers do not work systematically, being their corrections sometimes contradictory and ambiguous.

Recent advances in speech technology have allowed to provide individualized and automatic CF in CAPT tools. Care must be taken to design and adapt CF to specific pronunciation training activities with this technology in order to avoid erroneous feedback as false alarms and false accepts.

In this thesis, different types of CF are integrated into the experiments carried out to analyze which ones are the most suitable for building an effective CAPT tool. The aim is to automatically offer an appropriate training activity after a learner's erroneous input and a set of advice to overcome the specific pronunciation problem.

In this chapter, an overview about CF in SLA, from its essentials and different types to the possible integration into CAPT systems is presented. In particular, a revision of the literature about current CF techniques in CAPT experiments is carried out. First, the most significant features of relevant CAPT experiments with simple or isolated CF exercises are reviewed. Finally, other CAPT software systems with more advanced CF techniques are detailed.

4.1 Fundamentals of Corrective Feedback

In cybernetics, the term feedback refers to the process by which the system decides the next step after a previous action [108]. Following this definition, language learning systems must be similar to experienced teachers, whose working methodologies are adapted according to the needs of the learners [109], [110].

The preferable characteristics of CF in L2 learning are [111]: unambiguous, understandable, detectable, short, and should preferably take account of learner characteristics, both proficiency and literacy level. CF in pronunciation training of SLA can be divided into two main groups, *implicit* and *explicit*, in terms of whether or not the learner is informed of the corrected form of the error [112], [113]. In particular, implicit CF involves a repetition or clarification request after learner's erroneous utterance. According to [112], [113], there are three types of implicit CF:

1. **Conversation recast**: reformulation of a learner's utterance when a communication/connection problem arises.
2. **Repetition**: a prompt requests a new utterance attempt without pointing out the error.
3. **Clarification request**: a prompt requests the repetition of an specific utterance or part of it indicating that it has not been understood.

On the other hand, explicit CF is divided into six categories depending on whether the correct form is provided or withheld [112], [113]:

1. **Didactic recast**: reformulation of a learner's utterance even though no communication problem has been caused.
2. **Explicit correction**: direct signal that an error has been committed and its correct form is provided.
3. **Explicit correction with metalinguistic explanation**: the same as the previous one with the addition of a metalinguistic comment.
4. **Metalinguistic clue**: short questions or comments eliciting a correction from the student.
5. **Elicitation**: a prompt that verbally elicits the correct form from the learner.
6. **Paralingusitic signal**: a prompt that non-verbally elicits the correct form from the learner.

In LL, and in particular, L2 pronunciation training, providing phone-level CF with articulatory movements (i.e., manner and place of articulation [114]) to make students realize and correct their pronunciation mistakes, is a common followed method. However, it is not as simple as it appears and it causes difficulty to learners, particularly to beginners [115].

ASR-based CAPT systems offer an immediate and individualized feedback, a more intensive practice, opportunities for repair, and an anxiety-free context (autonomous practice), among others [65]. This is not surprising considering that ASR-based CALL systems can offer extra learning time and material, specific feedback on individual errors, and the possibility for self-paced practice in a private and stress-free environment.

However, there is also concern that ASR must be carefully integrated into CAPT systems since it can led learners to believe their pronunciation is accurate when it is not (false accepts), or that it is inaccurate, when it is clearly intelligible (false alarms) [37], [57].

TTS technology must be also considered an appropriate feedback resource in pronunciation training, particularly when combined with efficient teaching techniques. It can generate model performances of particular words and sentences (consequently, minimal pairs).

4.2 Corrective Feedback in CAPT Experiments

There are several factors that affect the selection of CF methods in CAPT systems, such as the training exercise type, the difficulty level, and the technology integrated. First, an easy approach is followed by CAPT systems in purely perception activities (see Section 2.1) in order to improve pronunciation. It consists in providing an assessment feedback with metalinguistic signals, without hints for improvement. On the one hand, in [82], [83], [37], and [85], a right/wrong sound of a chime/buzzer and a green/red color text is presented after a user's answer. On the other hand, another conventional way of assessment is to provide answer keys in slides [78].

Second, in production activities of most studies the teachers are the ones who provide feedback to the students after utterance mistakes with the computer, which entails a waste of time as they are not able to assist to all the students at the same time, resulting in lower efficiency. Several studies also urge the students to produce an audio sample similar to the native and to listen to both [87], [88], [89], [90]. In the study reported by [97] students are also able to listen to synthetic phrases but are not offered the possibility of recording themselves and compare both sounds.

Other CAPT systems show visual feedback as an enhanced spectrogram, in which students are able to notice a gap in their production which was something they had not been able to do through the imitative–intuitive approach alone [60], [89]. Recent techniques permit to transform the non-native spectrogram input to a spectrogram with properties of self-imitating feedback [99]. Another systems use computerized visual displays of pitch contours [91], [92], [93], [94], or novel techniques, such as flashing lights that show how much pitch variation the speaker has produced [95]. Finally, other works base their feedback on the scores or corrections made by other participants of the same experiment to the recorded audios [68], [69].

In the category of ASR-based CAPT systems, there are different approaches. On the one hand, some systems use ASR with very simple explicit feedback: HMM scores as feedback [60], [67], [103], right/wrong answers [65], or green (acceptable) / red (unacceptable) answers with words presented in appropriate contexts through audio and text and associated with representative images in case of mispronounced words [107]. In the case of mispronouncing sentences, the system offers to repeat the single mispronounced words [116] or the whole sentence [103], [117].

On the other hand, there are more sophisticated systems. The *Nuance Dragon Dictation* software used in [66] provides immediate written visual feedback to students via an orthographic representation of their result after each production attempt. Learners' goal is to produce each word and phrase correctly in a maximum time of one minute.

In [62], the *PLASER* software displays word's English spelling, its Chinese translation, a representative figure, and a pronunciation video-clip of a native American English speaker. It is also possible to listen to a word pronounced by a native speaker as many times as the user wants. As a result of producing an utterance, a 3-color feedback scheme for a GOP-based phoneme confidence score is displayed in the visual interface.

In [13], an ASR-based CAPT system provides explicit feedback on the screen. First, a theoretical video is presented, and then some related exercises. In each one, the orthographical transcription of the utterance pronounced by the learner is shown, together with a smiley-face emoticon and a short written comment. If a phoneme is wrongly pronounced, a sad-face emoticon appears with a red and underlined text, and a prompt to repeat. There is a maximum of three attempts per word in order not to discourage students.

In [118], the ASR-based CAPT system DISCO provides feedback in two contexts, in the remedial exercises and during the dialogues. Besides, learners can choose either a very explicit or a more implicit, communicative feedback strategy. First, the pronunciation errors are highlighted, allowing learners to immediately correct themselves. Second, recasts are provided in the conversation environment, repeating the student's response without the errors highlighting the erroneous graphemes, morphemes or words.

In [63], [64], the *Talk to me* software shows the articulation of the sounds. Photographic illustrations, music, and video-clips are provided in the form of animations. A score for the production is given, and if the program founds particular difficulties recognizing a specific word in the phrase, that word is highlighted in the text screen. The speech can be slowed down and the difficulty level of the speech recognition can be adjusted.

In [61], the *Fluency* system identifies production mistakes automatically, and offers text suggestions and hints for correctly pronouncing the phonemes targeted. This software also allows users to listen to their own records and a native speaker's ones. Besides, it lets individuals to read instructions on how to pronounce a sound; to see both sides headcut and front view, of the lips; and to listen to isolated words or in minimal pairs.

In [103], the *SPIRE-fluent* application provides an automatic feedback with visual scores (filled mugs) for learner's pronunciation quality in two contexts: for each word in a sentence and for the entire sentence. Besides, this system displays pauses and syllables present in the student's and expert's utterances for the given stimuli.

Despite the wide range of CF techniques applied to CAPT experiments described above, in [13] is reported that many of these CF strategies may be insufficient (i.e., right/wrong answer), inefficient (i.e., getting stuck trying to imitate a sound that you cannot imitate well), and may not be clear for the majority of learners (i.e., flashing lights that involve a lot of analysis time by the users [95]).

As a summary of this section, Table 4.1 shows a comparison of the individual CAPT experiments mentioned in the the state-of-the-art of this thesis about how they manage the aspects considered in the text with the four dimensions related to the training methodology: activities, ASR and TTS technologies, pronunciation assessment, and corrective feedback.

Reference	Activities	ASR/TTS	Assessment	Feedback
Bradlow *et al.* 1997 [82]	Minimal pairs: HVPT, Identification, Oral reading	-	Pre/Post (natives)	Right/Wrong, repetition
Guilloteau 1997 [86]	Words: Oral reading	-	Pre/Post (natives)	Right/Wrong, repetition
Akahane-Y. *et al.* 1998 [60]	Minimal pairs: Identification, Oral reading	ASR	Pre/Post (natives)	Spectrograms
Tomokiyo *et al.* 2000 [61]	Words & sentences: increasing difficulty, Oral reading	ASR	Pre/Post (natives)	Articulatory animation, native and own utterances exposure, external mirror
Wang 2002 [83]	Minimal pairs, HVPT, Identification, Oral reading	TTS	Pre/Post native rater	Right/Wrong, repetition, TTS exposure
Hincks 2003 [63]	Dialogues (oral reading of restricted utterance), custom difficulty	ASR	Pre/Post *PhonePass*	Articulatory animations, music, videos, utterance's score, wave form, pitch curve, slower speed
Mak *et al.* 2003 [62]	Words & Minimal pairs: Identification, Oral reading	ASR	Pre/Post (natives)	Videos, Bad/fair/good answer
Weinberg *et al.* 2003 [87]	Sentences: Oral reading	-	Pre/Post (natives)	Radiocassetes, videos
Hardison 2004 [91]	Sentences: Oral reading (pitch & intonation)	-	Pre/Post (natives)	Pitch contours
Hirata 2004 [93]	Minimal pairs, triplets: Identification, Oral reading	-		Spectrograms, Prosody graphs, native and own utterances exposure
Hardison 2005 *et al.* [92]	Sentences: Oral reading (pitch & intonation)	-	Pre/Post (natives)	Pitch contours, videos
Hincks 2005 [64]	Sentences: Oral reading (pitch & intonation)	ASR	Pre/Post (natives)	Articulatory animations, music, videos, utterance's score, wave form, pitch curve, slower speed
Lord 2005 [88]	Sentences: Oral reading	-	Pre/Post (natives)	Self-comparison with native speech
Neri *et al.* 2006 [57]	Words and Sentences: Identification, Oral reading	ASR	Audio listening (natives)	Text hypotheses
Burleson 2007 [100]	Minimal pairs and Sentences: Identification, Oral reading	-	Pre/Post (natives)	Right/Wrong, Score
Fangzhi 2008 [79]	Words and Sentences: Identification, Oral reading	-	Pre/Post (natives)	Right/Wrong, repetition
Lord 2008 [68]	Dialogues and Sentences: Oral reading	-	Pre/Post-test raters: participants	Self-comparison with native speech, participants' feedback
Neri *et al.* 2008a [13]	Words and sentences: Oral reading	ASR	Pre/Post-test (natives)	Right/Wrong phoneme level feedback, native and own utterances exposure, videos
Neri *et al.* 2008b [65]	Sentences: Oral reading	ASR	Pre/Post-test (natives)	Waveforms, Right/Wrong, native and own utterances exposure
Pakhomov *et al.* 2008 [107]	Words: Oral reading	ASR	Pre/Post-test natives/ASR	Right/Wrong, repetition
Hincks & Edlund 2009 [95]	Oral presentation	-	Pre/Post-test (natives)	Flashing lights, native and own utterances exposure
Lee 2009 [96]	Words: Identification, Oral reading	-	Pre/Post-test (natives)	Right/Wrong, repetition
Tanner & Landon 2009 [78]	Spontaneous speech, Identification, Oral reading	-	Pre/Post-test (natives)	Self-assessment (explicit answers in slides)
Alastuey 2010 [69]	Sentences: Oral reading	-	Pre/Post-test (natives)	Async expert's and participants' feedback
Pearson *et al.* 2011 [89]	Words and Sentences: Identification, Oral reading	-	Pre/Post-test (natives)	Spectrograms
Soler Urzúa 2011 [74]	Words: Oral reading	TTS	Pre/Post-test (natives)	TTS exposure, Teacher's feedback
Thomson 2011 [37]	Words: HVPT, Oral reading	-	Pre/Post-test (natives)	Right/Wrong, repetition
Chun *et al.* 2012 [94]	Words: Oral reading (pitch & intonation)	-	Pre/Post-test (natives)	Spectrograms and native audios exposure
Strik *et al.* 2012 [118]	Dialogues and sentences: Oral reading	ASR	Audio listening (teachers)	Text hypotheses and recasts
Thomson 2012 [85]	Words: HVPT, Oral reading	-	Pre/Post-test (natives)	Pitch contours, Right/Wrong, repetition
Kataoka *et al.* 2015 [71]	Sentences: Oral reading	TTS	Pre/Post-test (natives)	TTS exposure and own utterances
Liakin *et al.* 2015 [66]	Words and Sentences: Identification, Oral reading	ASR	Pre/Post-test (natives)	ASR - Text hypotheses, NonASR - teacher's recasts and repetitions
Eksi & Yesilcinar 2016 [101]	Oral presentations	TTS	Pre/Post-test (natives)	TTS exposure
Jayakumar *et al.* 2016 [116]	Sentences: Oral reading	ASR	-	Right/Wrong-words repetition
Luo 2016 [98]	Words, Minimal pairs and Sentences: Record-Listen-Compare	-	Pre/Post-test (natives)	Async teacher's and participants' feedback
Liakin *et al.* 2017 [97]	Sentences: Listen–rank, Listen–categorize, Listen–repeat.	TTS	Pre/Post-test (natives)	TTS exposure
Cheng 2018 [67]	Words & Sentences: Oral reading	ASR	Pre/Post-test natives/ASR	Text score, videos, native and own utterances exposure
Shinohara *et al.* 2018 [84]	Words & Minimal pairs: Identification, Discrimination, Oral reading	ASR	Pre/Post-test (natives)	Right/Wrong, repetition
Yang & Chung 2019 [99]	Words: Oral reading	-	Pre/Post-spectrogram manipulation (perceptual)	Self-imitating feedback: transformed spectrogram
Yarra *et al.* 2019 [103]	Sentences: Oral reading	ASR	-	Words' score, native and own utterances exposure, syllables and pauses

TABLE 4.1: Comparison of CAPT experiments in the literature.

4.3 Summary

Corrective feedback is one of the most important elements of a CAPT system to be successful in the learning process. Depending on the information requested and showed to the learner, CF can be implicit or explicit, with several sub-types of each one of them. In this chapter, a revision of the literature about CF techniques applied to CAPT systems has been conducted in depth. These techniques have been described and pointed out in the experiments revised from the literature. One contribution of this thesis is to automatically provide a set of CF techniques applied to a mobile CAPT system which guides users to achieve better pronunciation results, based on learner's choices and results.

Finally, an ASR-based CAPT system can offer an immediate and individualized feedback and a TTS system can generate pronunciation models, being both of them useful in the process of helping students to improve their perception and production skills. In this chapter, a comparison of the CAPT systems which include these technologies has been carried out.

Chapter 5

Game-based Learning with CAPT Systems

The possibilities offered by e-learning have been expanded in the last decade due to their convenience, flexibility, and promotion of active and independent learning. Self-motivation of students is key for this learning system to be effective and successful. Besides, the popularization of smart devices and the advances of technology have contributed to the expansion of online services and applications through the Internet for supporting L2 learning. However, these online courses and applications tend to register high abandonment rates after their first uses.

Gamification elements can be included into CAPT systems in order to motivate, engage, and stimulate user's experience. At the same time, CAPT systems can also be integrated into learning applications whose purpose is to improve L2 pronunciation. Social versions of learning games encourage a competitive and collaborative participation of individuals and groups. Furthermore, social game play is becoming more popular than individualistic game play, and literature suggests it is beneficial for players if it is designed correctly. Although this type of games constitutes a resource with great learning potential, there have been few attempts to empirically validate their inclusion in contexts related to pronunciation teaching with smart devices.

In this thesis, different game-based learning structures intended for CAPT-based learning games with gamification elements are proposed, aiming at the autonomous and social training of L2 pronunciation.

In this chapter, firstly an overview about learning digital games for language learning is presented. Next, the advantages and disadvantages of social (competitive and collaborative) games are explained with examples of state-of-the-art experiments. Finally, the most relevant gamification elements included in pronunciation training applications and studies are described.

5.1 Game-based Learning

One of the main reasons for using games for learning is that they motivate and encourage individuals to keep on training [119]. Their methodological approaches engage players to achieve and keep an intense capacity of concentration, pleasantness, and persistence, while challenges are closely adjusted to ability [120]. Besides,

a well-developed game goes beyond classroom boundaries and provides an incentive for social interaction [3].

Over the past decade, games have been included into higher education research tools [2], mostly due to their effective potential, including their entertainment value when teaching a certain skill [121]. Games also contribute to build productive social practices, helping individuals to take part in learning communities [3], [122].

As pointed out by [123], learning games can be categorized as **individualistic** or **social**, depending on how players organize their efforts. The former refers to users who play alone with the system, ensuring their own learning meets a preset criterion, independently from other participants. On the other hand, social games include not only the interaction with the machine but also with other players. Social practice can be classified into collaboration (cooperation among partners to accomplish shared learning goals and maximize their own and their teammates' achievement), competition (among competitors, trying to perform faster and more accurately than other participants), or a combination of both of them. In any case, a learning game ought to have at least the next three main characteristics according to [124]:

1. **A goal**. The mixture of objectives and events needed to finish the game. It must be precisely and clearly established. It is the most relevant aspect in the game since its success depends on it. Achieving a certain amount of points or number of badges could be examples of game goals.

2. **Obstacles**. Challenges and adversities intended to complicate the game in order to avoid triviality. For instance, limiting the number of times an exercise can be performed or enhancing the activity's difficulty along time.

3. **Competition or collaboration**. Players can compete against other users or try to beat the game itself. For instance, players can realize their individualistic game scores in a common leaderboard shared with other learners (individualistic efforts in an implicit competition). Other possibility is to promote challenges among users and divide the points depending on the results (explicit competition). Users can also form groups and try to reach a shared goal together (collaboration).

Social learning structures must follow certain typical aspects in order to be successful. Even though both cooperative and competitive learning strategies have common features, they can be clearly differentiated. In the case of competitive-learning scenarios there must be present at least six characteristics [123]:

1. **Negative goal interdependence**. If a user wins, the others must lose.

2. **Perceived scarcity**. Only the best players can reach rewards and achievements since their quantity is limited.

3. **Interaction with other parties**. It can be direct (with oppositional actions), indirect (parallel or sequential actions, turns), or nonexistent (i.e., playing individually to reach a final score). In this thesis, the concept of ***explicit competition*** is related to a turn-based indirect interaction with other subjects via challenges; whereas ***implicit competition*** refers to a nonexistent interaction among users in a common competition.

4. **Quantity of winners**. It varies from one, to few or many winners.

5. **Comparability among participants**. User's performances generate public information that can be optionally reviewed by the rest of the competitors.
6. **Winning rules**. The criteria for determining the winner must be clear. It can be objective or subjective, depending on the tasks.

On the other hand, a cooperative structure must take into account other six particular characteristics [123]:

1. **Positive goal interdependence**. Players must realize they can attain their goals if and only if their teammates attain theirs. It can be enhanced with positive reward interdependence (i.e., group rewards).
2. **Individual accountability**. Related to the individual share of the work by each teammate. Players must know the results achieved by the rest any time.
3. **Intergroup cooperation**. Groups can help others to finish the task successfully or compare their strategies and results.
4. **Desired behaviors**. Related to the specific teamwork and taskwork skills to be learnt by the players.
5. **Learning task**. Two aspects must be clarified: what and how must be completed the assignment/goal.
6. **Criteria for success**. Both the learning task and the level of performance must be cleared settled.

5.1.1 Social Learning Games

Nowadays the interest in social learning games is receiving a great attention from the literature. Social learning contexts for improving learning skills can be established with digital games, providing a learning scenario that offers learning contents, and a community that allows the condition for social learning [125].

Although there is not consensus in the literature about which approach is best for social learning games, the current trend in the state-of-the-art is mainly focused on collaboration over competition. While there are some studies that mention competition is related to violent and aggressive behaviors [126], there are also others that report similar effects in both approaches [127]. The effects derived from competitive learning games are influenced by other aspects of the game, such as the content, the rest of players, the complexity of the learning, or the game configuration, among others [128]. Actually, both alternatives are valid as long as they promote pro-social outcomes [129] and reduce as much as they can negative effects among players, such as aggression and aggression-related variables [130]. It is clear that more research is needed to determine under what conditions the competition can be most beneficial.

Particularly, there are some studies about L2 teaching that report the benefits of collaboration over the negative consequences of competition in games [5], [131]. Others experiments explain the implications, differences and advantages of both of them, collaboration and competition, on player's perception [132]. In the case of CAPT-based experiments, there are a few examples of collaborative learning, and to the best of our knowledge, there are no precedents in the case of competitive scenarios. For instance, in [133] three groups of students with individualistic and collaborative efforts are compared. It reports more qualitative gains and strategies

outcomes from the Collaborative CAPT Group. In [134], both the individualistic CALL and the collaborative computer-mediated technique approaches are reported positively by the students. However, motivation in games can be associated to the term "challenge", with positive outcomes [135]. Competitiveness, in the context of game-based learning, also helps trainees to achieve their learning goals and perceive higher ability skills [4].

There are several studies that report motivation and engagement enhancement due to competition [136], [137], [138]. In [139], the comparison between students' scores in a perception training controlled-competitive experiment tries to motivate them to improve their own results. In [140], an educational mathematics game shows the positive effect of competition in a collaborative learning situation for above–average students.

Individualistic and competitive approaches are also directly compared in some experiments of the literature. A high enjoyment, future play motivation, and high physical intensity are reported in [141], thanks to competitive scenarios since parallel competition in separate physical spaces overcomes individual gaming. In [142], positive experiences when a competitive context is provided to competitive individuals are described; whereas detrimental effects are reported for the less competitive ones. In [143], it is reported that the competitive configuration in learning approaches requires further research since better results (but not statistically significant) were achieved by the users of the non-competitive condition in comparison to the competitive one. In [132], the players of a cooperative setting experienced greater enjoyment than those in a competitive configuration.

5.2 Gamification and L2 Pronunciation Training

Gamification is defined as the use of game design elements in non-game contexts so as to enhance participant engagement and encourage desired behaviors with a product or service [144]. Gamification is also intended to reduce abandonment by designing attractive applications that generate pleasant and beneficial affection [145].

In the particular field of game-based methods and strategies for learning contexts, gamification uses game-based mechanics, aesthetics, and game thinking to engage individuals, promote learning, and solve problems [146]. Besides, educational gamification helps individuals to be immersed in learning, improves their motivation, and brings them playfulness [147].

The first commercial educational language learning tools, such as Sanako[1] and Rosetta Stone[2] appeared in the 1990s. It is not until the second decade of the twentieth century when the modern online applications appear (mobile, web and desktop), such as Duolingo[3], Busuu[4], or Babbel[5]. The former mentioned group of tools are pronunciation training services with methods and strategies based on content choice, focused on either self-training or academic institution solutions. However,

[1] http://www.sanako.com/
[2] http://www.rosettastone.com
[3] http://www.duolingo.com/
[4] https://www.busuu.com
[5] https://babbel.com

the latter group of pronunciation training applications have changed the paradigm by including gamification elements with an evident intention to improve the user's learning experience [148].

Although the number of applications with gamification elements intended for L2 pronunciation training is increasing, there are scarce experimental studies that validate their effectiveness. For instance, [149] describes a card-based game for L2 vocabulary acquisition with speech technology. In [65] a word game based on stories for children with an ASR system, PARLING, is presented. In [150], the Polish language is taught as a user role-based game, in which its complex grammar system and lexical problems are presented to learners as tasks and activities. In [118], a CAPT game for practicing Dutch oral and grammar skills based on speaking practice and feedback is reported. There are other innovative ways of pronunciation teaching with gamification, such as a multi-language karaoke application, SLIONS [151] or a recursive dialogue game for a personalized pronunciation training [152]. The previously mentioned studies and applications share game design elements which can be classified into [153]:

1. **Points** are the most extended and basic elements in language learning tools. They consist on a numerical representation of the result of an activity performed by a learner. Their scale can vary from complex ranges of numbers to simple binary outputs reporting user's success or failure in the activity. In particular for pronunciation training, this important resource could be used to assess the goodness of the user interaction and to measure the user proficiency [150], [154], [155]. For instance, in [151], an overall score (0 to 100) from a user's karaoke performance is shown to the learner. Besides, points can involve the accomplishment of user's levels and badges and can lead to earning rewards in Duolingo and Babbel.

2. **Badges** display symbols or messages that represent user achievements. They are intended not only for informing user's about their performance but also to motivate them to keep on training [149], [151], [152]. For instance, in [150], "The Lord of Memory" badge is given to a student who remembered most of new words of previous classes. In Busuu, individuals earns several different badges after concluding courses or talking tests. In Duolingo, users can earn badges after completing 10, 50, and 100 lessons, or 5, 10, and 30 skills, in addition to an extra incentive for making progress with the lessons, among others.

3. **Leaderboards** show a ranking that permits users to compare their relative success as regards the performance of other players [153]. This competitive indicator of progress allows users to contrast their own level regarding other learners, contributing to assemble a self conscience of level. It is also interesting because it also permits to configure competitions as a mean to promote social interaction between users. This game element is commonly used in social language learning applications, and almost non-existent in the state-of-the-art about pronunciation training studies with CAPT. For instance, in Duolingo and Babbel, progress is measured in terms of gaming-like elements by gaining experience (XP points), and leveling up, which affect to their social leaderboards.

4. **Performance graphs** provide information about the player own progress over time. The difference with leaderboards is that, in this case, performance graphs do not compare the user's performance to other players. Thus, it is an individual reference indicator instead of a social one. In particular, this resource is relevant for both students and teachers, since it can trace all right and wrong interactions with the system and can lead to personalize user's training content based on the results achieved [152]. For instance, Sanako provides a complete dashboard for teachers to follow up the progress of the students in the language laboratory. In Duolingo and Babbel, graph statistics and historical records are available to users. In [118], a final report about all mistakes made by the learner is given after each conversation.

5. **Meaningful stories** stand for the narrative in which the gamified activities and characters are included in. They give a meaning far beyond the only purpose of achieving points and badges [146]. In particular, they could be used in language learning applications to connect different activities in the same flow of exercises. For instance, in [65], the word-based pronunciation training system, PARLING, displays well-known children's stories. In [156], personalized activities in significant 3-D virtual environments are shown. Duolingo offers activities related to particular stories which can change with game progress, presented by an owl character, Duo.

6. **Avatars** represent players in the game and allow them to communicate themselves as well as objects within the environment. They vary from simply approaches as pictograms, to more complex ones, such as three-dimensional (3-D) representations. They are intended to enrich the user's experience during games. They are widely used in language learning games in virtual environments. For instance, in [150] a warrior-based avatar represents each group of students in the game. In [156], 3-D human-based avatars can interact with the whole environment. In [157], human-based avatars permits also more than two people to be involved in the conversations.

7. **Teammates** (teamplayers) are the rest of players in the game. They could be real or virtual ones, and can lead to conflict, competition or collaboration [146]. In particular, L2 learning applications with gamification elements must help users to overcome not only their own communication barriers but also to compete with other players [150]. For instance, Duolingo encourages users to compete with their friends to see who learns faster. In [150], each learner has a role in the game, and they form small groups, affecting other player's actions. In [152], learners interact with virtual and simulated players in dialogues.

The impact of these game elements on autonomy, psychological needs of competence, and social relatedness is analyzed in [158]. Some of them, such as points and leaderboards can be considered as extrinsic incentives for promoting performance on image tag task [159]. The design of effective leaderboards based on individuals preferences is analyzed in [160], which concludes that competition is a media rather than purpose when there are leaderboards in gamified applications.

5.3 Summary

E-learning can be defined as any kind of learning or development content administered in a digital way. Recent technological advances have allowed researchers to integrate CAPT systems in online learning contexts, allowing users to learn anytime anywhere while keeping motivated. In this thesis several motivational elements are included in the CAPT prototypes of the experimentation. In particular, game approaches are carried out by means of prototypes of CAPT learning games for pronunciation training. Therefore, a large amount of data is automatically gathered susceptible to be in a speech corpus.

Part III

Experimental Procedure

Chapter 6

Experimental Framework

Developing an effective L2 CAPT tool for learners of a particular native L1 language requires the design of a proper set of training activities and corrective feedback techniques, along with a reproducible objective method of assessment, and an appropriate choice of speech technology. The purpose of this chapter is to present the dimensions of the experimental procedure, demonstrating an understanding and applicability of the specific concepts, theories, and technologies which are relevant and necessary for the experimentation. Firstly, the importance of minimal pairs in pronunciation training is described. A novel protocol for elaborating minimal pairs lists taking into account the speech technology integrated in the CAPT tool is also detailed. Secondly, the essentials of the training activities cycle included into the experimental prototypes are specified. Third, the strategies carried out to assess user's pronunciation improvement with the CAPT tools are presented. Fourth, the fundamental principles and preferable characteristics of ASR and TTS systems are described. The selection of some state-of-the-art ASR and TTS technologies is also motivated, including a general outline for building a personalized ASR included in the CAPT tools developed. Then, the main feedback strategies adopted for each one of the experiments are commented. Sixth, the gamification elements included in the experimentation are mentioned. Finally, the aspects taken into account to select participants to carry out all the experiments presented in this work are mentioned and which will be discussed in Chapter 7.

6.1 Minimal Pairs

After a review of the available methodological approaches to L2 teaching (see Chapter 2), the NCM has been partially followed as the basis for the design of learning activities in the CAPT systems we developed for the different experiments presented in this part of the thesis (see Section 2.1.1 for more details). As a reminder, the main aspect of this approach is the use of minimal pairs in a specific cycle of pronunciation activities, taking into account the learner's L1 and L2. The concepts and methods of these approaches have been extrapolated to the experimental work to several languages, such as English and Spanish.

Two different general approaches will be experimented for users' guidance. First, a gamified playing methodology with free selection of activities in a learning game. Second, a guided and controlled pedagogical methodology with recommended activities of feedback based on user's results that lead to a common end. Both approaches integrate current speech technology adapted and personalized to the CAPT systems.

The term *minimal pair* in this thesis refers to a pair of words that differ in only one sound, changing their significance completely [161]. This concept can be applied to most languages. From a pedagogical point of view, working with minimal pairs helps individuals to realize the potential risk of altering the meaning of a word by changing a single sound in an incorrect utterance. For instance, in American English, the minimal pair *Sue – zoo* contrasts the phoneme /s/ (voiceless alveolar fricative) and the phoneme /z/ (voiced alveolar fricative). Another example in Spanish could be *rey – ley*, which contrasts the phoneme /r/ (trill) and the phoneme /l/ (lateral). Furthermore, not only consonants can be compared, but also vowels. For instance, in Estonian the minimal pair *noor – nöör* contrasts the phoneme /o/ (close-mid back rounded vowel) and the phoneme /ø/ (close-mid front rounded vowel).

Although lists of paired words were originally used to extract phonological catalogs from relatively unknown languages, the minimal pairs technique is also being used today for the teaching of the pronunciation of L2 [42]. Minimal pairs are also adequate to elaborate activities of sound exposure for different pairs of phonemes, in order to familiarize learners with them. They also facilitate performing perception activities to discriminate between different sounds in words since minimal pairs test learner's ability to discriminate between the elements of the pair. Finally, minimal pairs also provide a reference to prepare pronunciation exercises which require production of the utterances of each word of the contrast [39].

The minimal pairs technique is applied in some of the state-of-the-art experiments for L2 exposure, perception, and production training (see Section 2.3.1 and Section 2.4.1). In some cases, the differentiation between words of minimal pairs is a challenge not only for L2 students, but also for current speech technologies, leading to erroneous feedback. These problems convey a careful selection of the set of words of each minimal pair, which will be discussed and presented in Section 6.1.2 and can be summarized as follows:

1. **Homophones**. Words that have the same pronunciation but different meanings, origins, or spelling. For instance, in American English, the words *heal – heel* are homophones.
2. **Out-of-vocabulary (out-of-context), OOV, words**. Unknown words that appear in the user's speech but not in the recognition vocabulary.
3. **Infrequent words**. Words which have not enough data to be correctly processed. They are usually discarded since they are more likely to be misrecognized by an ASR system or bad synthesized by a TTS one.
4. **Breached boundaries**. These problems appear in the context of two or more linked words. For instance, coalescence, linking mechanisms, accommodations, assimilations, elisions and weak forms are examples of them. In this thesis, the ASR technology does not face these problems since minimal pairs are isolated words (surrounded at both ends by silence).

6.1.1 Languages Covered

In this thesis, a specific battery of minimal pairs for each experiment has been elaborated and adapted to the L1 and L2 language of the learners, and classified by sound contrasts. Gathering these lists of minimal pairs has not been a trivial task since it has been necessary to count on experts knowledge about the specific pronunciation problems of the individuals, and to obtain a lexicon big enough as to

have relevant coverage of candidate words for the minimal pairs. For these reasons, a semi-automatic protocol has been designed to elaborate minimal pairs lists for any language (see Section 6.1.2).

In particular, the Castilian Spanish language (es_ES) has been the first one taken into account since it is defined as the L2 target of two research projects this thesis is related to (see Section 9.3.9) and Valladolid hosts every year thousands of students of Spanish as L2. Secondly, the American English language (en_US) has also been considered into some of the experiments due a 70.9% of the Spaniards study English as L2 (EFL) [162] and it is the reference language in the NCM. Finally, as a result of the different collaboration with other institutions and colleagues (see Section 9.3.8), five additional languages have also been integrated five languages to the experimentation: simplified Chinese – Mainland China (cn_ZH), European (pt_PT) and Brazilian (pt_BR) Portuguese, German (de_DE), and Estonian (et_EE).

6.1.2 Minimal Pairs Selection Protocol

One of the key contributions of this thesis is the design of a protocol for the semi-automatic elaboration of the minimal pair lists which are included into a CAPT tool (one list for each minimal pair contrast). These lists of words are indexed by phonemes of the target L2 language and they are suitable for their integration into CAPT systems with speech technology. The specific procedure discussed in this section requires the joint collaboration of engineers, scientists, linguistics, and phonetics experts. Figure 6.1 represents the steps to follow in our proposed protocol and they are described below.

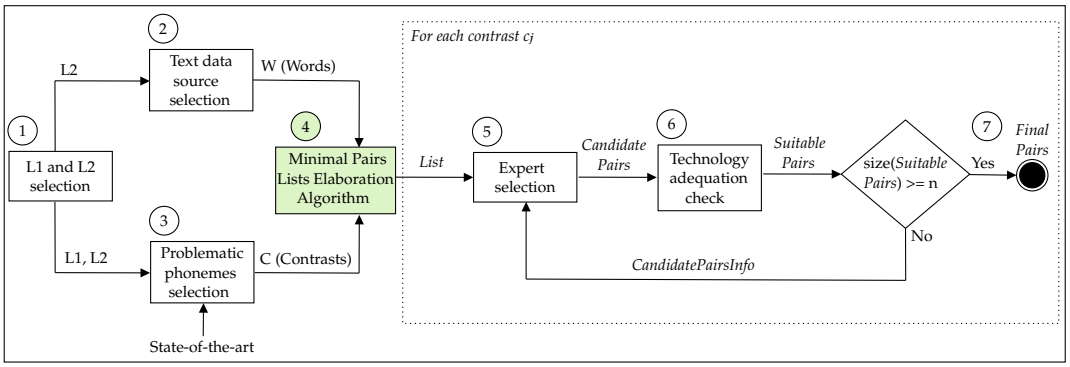

FIGURE 6.1: Minimal pairs lists selection protocol scheme.

Let *List* be a list of all possible minimal pairs of the studied contrasts c_j of an L2 language, L, extracted from a lexical corpus, defined as:

$$List = \left\{ P_{c_j} = \{(w_1, w_2)\} \,|w_1, w_2 \in L, \quad c_j \in Contrasts(L) \right\}. \tag{6.1}$$

where P_{c_j} is the set of all possible minimal pairs that consists of a pair of words w_1, w_2 of a specific contrast c_j of the set of words W extracted from the text source.

Let *CandidatePairs* be a subset of *List*, *SuitablePairs* be a subset of *CandidatePairs*, and *FinalPairs* be a subset of *SuitablePairs*, defined as:

$$FinalPairs \subset SuitablePairs \subset CandidatePairs \subset List. \tag{6.2}$$

Let $CandidatePairsInfo$ be the word information obtained for each minimal pair of the *CandidatePairs* list and n the minimum desired number of minimal pairs in each list. Seven steps of the protocol are defined as follows:

- **Step (1).** In this starting point the native and foreign languages are selected. These targets will determine the rest of steps, since the aim is to adapt L2 words to the phonetic characteristics of L1.
- **Step (2).** The text data sources are chosen according to the L2 and the word features needed (i.e., scientific texts, isolated words from dictionaries, literary novels...). This selection could also be done manually by experts, selecting words based on their personal experience.
- **Step (3).** A revision of the literature about the most complex phonemes of L2 according to the target L1 is carried out. A set of contrasts (C) is obtained to group all words (W) in the following steps. Each one of these contrasts consists of a pair of phonemes of L2. Both steps (2) and (3) can be performed in parallel.
- **Step (4).** An **algorithm for elaborating minimal pairs lists** which looks for the minimal pairs lists in the data sources of step (2) is executed (see specific details in Appendix A about the pseudocode, time and space complexity, and the input/output specification). The algorithm's output is a *List* with all possible minimal pairs of the contrast, c_j, with their phonetic transcription, ordered by their frequency of appearance, and indexed by the L2 phonemes of the contrast associated to the minimal pair (see Equation 6.1).
- **Step (5).** From this point, each list of minimal pairs (classified by phonemes) is analyzed individually. A list of *CandidatePairs* is obtained after an expert revision from the *List* of minimal pairs. This revision consists of, first, automatically filtering out the most frequent words from the text source, and second, discarding words that could be inappropriate according to the expert's experience and expertise (i.e., excessive length, disuse, or vulgar). The number of pairs selected must be equal or higher than n.
- **Step (6).** The adequacy of the words to the speech technology integrated into the CAPT tool is checked (see the description of possible problems in Section 6.1). The set of word pairs, P, that get the highest values in the expression described below are included in the output *SuitablePairs*.

$$SuitablePairs = \left\{ P_{c_j} | TTS(w) \geqslant \Psi \& ASR(w) \geqslant \Omega,\ \ P_{c_j} \in CandidatePairs,\ \ \forall w_i \in P_{c_j} \right\} \quad |SuitablePairs| \geqslant n \tag{6.3}$$

where Ψ and Ω stand for the minimum threshold of suitability of each speech technology system (see Equation 6.4 and Equation 6.5, respectively) for each word w of the pairs of the set P_{c_j}.

$TTS(\cdot)$ represents the *total TTS suitability* of words of the minimal pairs for the specific TTS engine integrated into the CAPT tool:

$$TTS(w) = Intell(w) - Complex(w) \tag{6.4}$$

where:

- $Intell(w)$ stands for an intelligible synthesized model of the word w by the TTS. A binary value is assigned by an expert when the word is synthesized. It depends on whether is intelligible for the expert (1) or not (0).
- $Complex(w)$ is the *complexity* of a word w of a minimal pair. This term in this special context stands for other sound difficulties of the word apart from the minimal pair phoneme targets. If this target is a vowel, the complexity usually refers to the other consonants of the word, and vice versa. It is intended to minimize side effects when listening to the synthesized word (i.e., difficult consonant aggregation before a contrasting vowel, due to its complexity). A binary value is assigned by an expert when the word is synthesized. It depends on whether there is complexity (1) or not (0).
- $TTS(w)$ is the quantitative *TTS suitability* of each word w of the minimal pair p, taking values in $\{-1, 0, 1\}$. The value assigned to Ψ in this thesis is $\Psi = 1$ (the maximum possible). This means that the synthesized words must be completely *intelligible* and without *complexity* since learners will face them in the training activities as corrective feedback. This value allowed us to find at least ten optimal minimal pairs for the activities which included TTS technology in all the experiments.

$ASR(\cdot)$ represents the *total ASR suitability* of the words of the minimal pairs for the specific ASR technology selected:

$$ASR(w) = \sum_{j=1}^{6} ReadSucc_j(w) - Transf(w) - Complex(w) \tag{6.5}$$

where:

- $ReadSucc_j(w)$ stands for the result of six readings of the word w by a native. If the ASR recognizes the word (or one homophone) in the first position of the results, a binary value is assigned (1–true, 0–false). The experts can find homophones in the results due to their L2 and phonetics knowledge. In this case, the homophone is associated to the word of the minimal pair for future uses of the ASR.
- $Transf(w)$ is the *transferred pronunciation* from an L1 speaker of an L2 minimal pair word reading. A binary value is assigned depending on whether or not (1 or 0, respectively) the ASR recognizes the word in the first position of the results even though it is produced as an L1 sound.
- $Complex(w)$ is the *complexity* of a word of the minimal pair in TTS.
- $ASR(w)$ is the quantitative *ASR suitability* assessment of a word w of a minimal pair p, taking values in $\{i \in Z | -2 < i < 6\}$. The value assigned to Ω in this thesis is $\Omega = 5$ (the penultimate maximum value possible). This value allowed us to find at least five optimal minimal pairs for the activities which included ASR technology in all the experiments.

- **Step (7).** If the size of the *SuitablePairs* list is equal or higher than the desired n length, the list is turned into the *FinalPairs* list of the minimal pair contrasts and can be integrated into the CAPT tool. Otherwise, the protocol

returns to step (5) with the information gathered in previous steps (5) and (6) ($CandidatePairsInfo$). The expert must repeat the selection of pairs taking into consideration the information obtained in such previous steps.

6.2 CAPT Methodology

It is generally accepted that students learn in different ways, according to their individual abilities [109]. On the one hand, some learners make a great leap from perceptive memory to precise production by just following their intuition. On the other hand, there are students who need more explicit instructions as feedback. In fact, nowadays there is still controversy in phonetics about the benefits of giving or not explicit instruction for pronunciation improvement [90].

Over the past ten years, the number of pronunciation teaching experiments and studies has greatly increased with diverse promising findings, not only with traditional instruction methods, but also with emerging technologies, as described in Chapter 2. In fact, these studies include explicit pronunciation instruction, which can be defined as the provision of articulatory (how to produce) or/and auditory (how to hear) information about L2 segmental and suprasegmental features, in order to measure its impact on L2 learner's pronunciation proficiency [163].

In particular, the design of the pre- and post-tests, the training protocols and their assessment are partially based on NCM. Other ideas that illustrate the differences between similar sounds in contrast lists, such as showing multimedia resources or designing a user-friendly interface in a software program, are inspired by [41], [42], [43].

The different activities for pronunciation training performed along the experiments based on the NCM presented in this thesis are described in the next subsections (see a comparison in Table B.2 in Appendix B). The fundamentals of these activities and how they could be integrated and combined into a specific methodology are included.

6.2.1 Explanatory (Theoretical) Activities

The first type of activities related to the presentation and explanation of the pronunciation concepts are short multimedia videos, two or three minutes long, which explain and auditory illustrate the articulation of the target sound of the minimal pair in the learner's L1. They consist of a read explanation prepared by an expert in both L1 and L2, using the same written information to be found later in the CAPT tool device's screen. It also includes animations of the sagittal section of the human vocal articulatory system and short videos taken in front of the face of a native speaker pronouncing the target sounds. The very nature of the videos allows individuals not only to listen to the video's explanations and see the animations, but also to repeat the sounds and word examples by themselves and at the same time.

These videos follow the NCM approach, that aims at providing not only a perceptive induction–oriented experience but also a deductive one with NCM–based instructions, which point out the kind of transformations which must be practiced upon an L1 sound in order to turn it into a close one in L2. The wording in the videos is intentionally redundant: the same instructions are usually expressed once in simple technical terms, and then in a friendlier, more impressionistic and intuitive

terms —*'Pronounce Spanish {e}, and now try to give it a little bit of {a} flavor'*. Both articulation and perception cues are used, in an attempt to address different learning styles. The IPA alphabetic system of phonetic notation is used to textually represent sounds in the videos and in the rest of activities. It is assumed that any particular aural memory benefits learners in terms of recollection from attachment to a particular non-ambiguous visual form.

A second type of explanatory activities, not yet addressed in our experiments, corresponds to short tips or advises that appear when a user mispronounces a word. They are defined by experts and must be easy to understand by learners, being offered to them in their L1. For instance, when a Japanese speaker produces wrongly the sound /θ/, a prompt from the system appears with the tip *'Put your tongue between your teeth'* or with the tip *'The air must come out through the center of your mouth'*.

6.2.2 Exposure Activities

They consist in a cycle of listen–repeat–compare tasks with minimal pairs. In particular, subjects listen to the words of a minimal pairs list a limited number of times (mandatory listening) and try to imitate their sounds [15]. An inductive discovery of the L2 phonemes from a first-hand perceptive experience is achieved in order to assist their assimilation [161]. Individuals become familiar with the distinctive phonemes within such sequences of minimal pairs, randomly presented (listen). The aural correlate of each word of the minimal pair is played a maximum of five times by a TTS system, each repetition being noticeably slower than the previous one. Finally, learners must record their own realization of the words —at least one time— (repeat) to compare it with the synthesized versions, by listening to both sounds (compare) as many times as they want (requested listening).

6.2.3 Discrimination (Perception) Activities

This type of exercises lets users to test their ability to discriminate between the elements of minimal pairs [164]. Identification and recognition success of L2 phonemes is achieved in this stage [36]. In this thesis, learners listen to the synthesized aural correlate of any of the words in each pair (mandatory listening) and must match it with the correct written form given on the CAPT system's screen (identification task of perception, see Section 2.1). One attempt of word selection is the sequence allowed per minimal pair. Users can listen to the word as many times as they want (requested listening).

6.2.4 Production Activities

These training tasks aim at helping users to alternately produce the sounds of a minimal pair by accommodating to a mental representation of them (previously acquired in earlier stages) [161]. They try to rematerialize (produce) the mentally acquired phonemes since they are no longer imitating an externally presented model [164]. Thus, the possible differences between mental and physical forms can be detected by the learners; being possible to notice their own errors [113]. They are also expected to self-diagnose accuracy, and know when self-correction is available.

Learners must separately read aloud[1] both words of each minimal pair. The ASR system supports the CAPT tool by providing feedback. In particular, the supplied

[1] While we eventually record the utterance for further off-line processing, if any.

prediction of the ASR system is composed of a list of *n*–best possible text hypotheses (*n* is adjusted in each experiment), ordered from highest to lowest confidence rates. In our case, these hypotheses are words and each possible word is followed by a numeric value (*g*–score), which is proportional to the reliability of the prediction (from 0% to 100% in a scale [0, 1]) although there is no documentation available on the specific meaning or interpretation of this score as a likelihood or similar. Thus, the utterance is considered correct as long as it is within the list of *n* elements returned by the ASR. For instance, an ideal utterance of the word *mass*, produced by a native American speaker would be associated to a *g*–score = 1.0 and a *5*–list of strings as follows: *"mass"*, *"Mass"*, *"masse"*, *"masts"*, *"mass."*. The sequence of maximum wrong production attempts per word is adjusted in each experiment.

Furthermore, synthesized models of the words are available to learners, who can play them as many times as they need while in production activities (requested listening) in order to improve their self-perception of the correct pronunciation to be obtained. The maximum number of listening attempts per word is adjusted in each experiment.

6.2.5 Mixed Activities

This type of exercises consists of mixing up both, perception and production activities, being intrinsically more difficult. The strength of the relationship between them may vary according to the proficiency of the speaker-listener and the target (L2) sounds [36]. In particular, in the mixed activities both kind of activities (discrimination and production) are sequentially interleaved. While in each one of isolated discrimination and/or production activities users can fully concentrate on these tasks individually, the mixed activities represent a situation closer to real communication, where readiness both to understand and produce language must coalesce. From a methodological point of view, they convey an extra difficulty and are included it as a way to review and test both modes at once as well as the ability to shift between them, simulating a real conversation.

6.2.6 Selection of Activities

A different set of activities is presented in each experiment depending on several factors: activities freedom of choice, activity goal, and progression along time. In terms of freedom to select activities, they can be restricted by the system (guided-based) or freely selected by users (free will). The former case refers to activities that belong to a previously defined by experts controlled protocol that leads to a common end. In particular, the system recommends a specific type of activity based on user's results. On the other hand, educational game-based prototypes of the experiments presented in this thesis give learners the freedom to choose the pronunciation activities. In some cases, the number of times the activities can be performed is limited.

Activity goal includes either training or playing. An specific target sound and any type of activity can be chosen in the first case in order to let learners to train at their own peace. However, in playing activities in which there is a final reward and participates other subjects, these elements are not possible to be selected.

The last aspect to take into account is progression over time. The content and difficulty of the activities change (locking/unlocking and decreasing/increasing, respectively) according to the evolution of user's results and success rate values.

6.3 Assessment

The ultimate goal of pronunciation assessment is to generate a score for a non-native speech utterance automatically and obtain results comparable to a human teacher/expert. As pointed out in Chapter 3, there is not a standard protocol for assessing user's pronunciation improvement with mobile CAPT tools and speech technology. From a realistic and experimental point of view, two kinds of assessment strategies can be combined, which are usually classified into objective and subjective categories. The goal of a mixed approach is to help teachers and students to save time and resources by obtaining an objective score about learner's performance with the system. The next subsections describe the potential data sources used in each experiment for the assessment of user's pronunciation improvement (see a comparison in Table B.3 in Appendix B).

6.3.1 Subjective Assessment

In the experiments carried out in this thesis, three different subjective approaches have been applied in which both students and educators take part (see Section 3.1 for more details and references):

- **Perceptual tests**. Pronunciation at segmental level quality can be characterized by the perceptual parameters identified by human experts, while evaluation is the distance between the target and the reference phoneme characteristics. Giving scores can be either written manually or typed with a computer via a personalized web page. Raters must apply the same rules to score the utterances. These tests can be carried out at the beginning (pre-test), at the middle (middle-test), at the end (post-test), and some days or months after the experiment (delayed post-test).
- **Questionnaires**. They provide quantitative data from learners during different stages of each experiment (i.e., pre-quest and post-quest). The users' demographics, their opinion about the system's experience (UX), the grade of motivation, and the reasons for participating or abandoning in the experiment are examples of topics included in these questionnaires.
- **Focus groups**. This qualitative technique is carried out at the end of the experiment. A set of predetermined questions is asked to participants in a planned discussion, while the moderators take notes. Also, the whole session is recorded via audio or video with the participants' consent. Individuals can be classified by common features into different focus group sessions. Apart from questionnaires' information, focus groups allow to obtain extra relevant information in an alternative way.

6.3.2 Objective Assessment

Giving an objective assessment about user's performance helps learners and educators to keep track of the evolution of pronunciation improvement along time while saving time and resources. However, technology must be specifically adapted to the task since false alarms and false accepts can appear. Also, objective scores should be correlated to expert human scores in order to ensure their validity.

The CAPT systems developed in this thesis resort to ASR to obtain binary (right or wrong) ratings of word pronunciation. In this thesis, two instruments are employed to provide objective assessment (see Section 3.2 for more details and references):

- **Objective tests**. User's utterances of pre/middle/post/delayed post-tests can be evaluated not only by experts but also by speech recognition. It must be specified the objective of the evaluation (i.e., the whole utterance, a specific syllable, or the target phoneme).
- **Game scores**. User's interaction with the system is kept into log files. This information includes user's activity results and several **metrics** can be defined. Learners realize about their performance and the system can personalize the activities content in function of this assessment. For instance, in discrimination activities, a right/wrong answer with numerical score is provided to the user and in the production ones is also shown a message with the most probable text spoken by the learner, evaluated by an ASR. Also, a *game score* at different stages can be correlated to the subjective and objective scores of the perceptual tests.

6.4 Speech and Software Technologies

6.4.1 Automatic Speech Recognition

Automatic speech recognition (ASR) is the use of computer hardware and software-based techniques to identify and process human voice [165]. It is also known as automatic voice recognition (AVR), voice-to-text, speech-to-text, or simply speech recognition. ASR is not a simple task since there are different factors that affect it directly, such as the age and accent of the speaker, the codec used for the audio and compression artefacts, the sample rate, the background noise, the length of silences, the reverberation from varying the acoustic environment, and the artefacts from the hardware. As shown in Figure 6.2, ASR converts the user's speech input (O) into a sequence of text hypotheses (W).

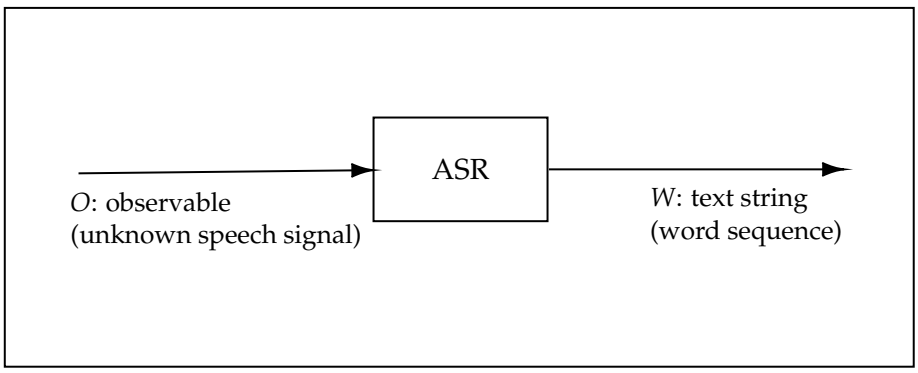

FIGURE 6.2: Conceptual ASR system.

The basic functionality of ASR is similar to any pattern recognition system, that is, some models are trained to subsequently recognize speech [166]. Two main model types can be distinguished:

- **Acoustic models** represent the relationship between an audio signal and the phonemes (or other linguistic units). That is, they are the statistical representations of a phoneme's acoustic information. In most cases they can be considered as task-independent models.
- **Language models** assign probabilities to combining acoustic models in order to form sentences (list of words). Generally, they involve task-dependency.

A **lexicon** (pronunciation model) serves as the link connecting the acoustic models and the language models. It is typically handcrafted by expert humans (a costly and time-consuming process).

The uses of ASR are not limited to CALL. There are several practical applications of ASR systems such as to identify the words a person has spoken (i.e., dictations, voice commands, etc.), to provide information and to forward telephone calls, to help people with disabilities (i.e., fluidity or transmission of a conversation to a person with hearing problems) and to authenticate the identity of the person speaking into the system. In this thesis, ASR technology is included as an element of assessment and feedback for speech pronunciation.

Choosing between an existing commercial off-the-shelf ASR system or a custom made and personalized one, is a crucial decision to obtain better or worse results in the specific domain of the problem to resolve. Sometimes, commercial ASR systems, such as Google ASR, Nuance Dragon, or Human VoiceBase are suitable for specific tasks, such as voice commands, telephone calls, or dictation assignments. However, a free and open-source ASR system as Kaldi could be better for specific educational purposes. In this thesis, both ASR types are incorporated in the experiments. Several features have been taken into account in order to select the ASR solution to be included in our CAPT tools [165]:

- **Accuracy**: testing the ASR system with experts for its suitability before experimenting.
- **Confidence measures**. The majority of ASR systems provide scores produced by extracting confidence features from the computation of hypotheses at the phonetic, word, and utterance level. Then, these features are processed using an accept/reject classifier for these hypotheses. They can be combined to linguistic scores and pragmatic constraints to offer to the speaker a corrective feedback.
- **Continuity**: determines whether the system can recognize continuous speech or a pause between word and word must be forced.
- **Custom vocabulary**. The accuracy will be higher if the target set of words is closed.
- **Documentation**. Knowing the possibilities offered by the system would help to personalize and reach the objectives.
- **Environment robustness**. Audio noise, stress, and sample rate are the most common factors that affect ASR systems performance.
- **Languages**. Each language, accents, and dialect variants must be trained separately. It depends on the unit used to build the ASR models. For instance, languages with common phonemes can share some models if they are phoneme-based.

- **Learning curve**: difficulty of understanding the ASR system design to integrate and personalize it (time and resources).
- **Price**. The current trend in commercial ASR systems is to pay in transactions per unit of time instead of buying a complete ASR system. Other possibility is to use a reduced version of their capabilities for a limited period of time. Open-source systems are for free.
- **Word level timing**: enables accurate linking to audio segments and helps enable comparison/merging of transcripts from multiple sources (i.e., taking punctuation from one transcript and applying it to another).

Google ASR Technology

Google's speech recognition[2] is a general-purpose and commercial off-the-shelf service available for more than 120 languages and variants. It combines the power of cloud-based computing with the latest technology. Besides, thanks to the data gathered from millions of users using all software applications of the company, Google have improved the accuracy of their machine learning algorithms for achieving better results. To the best of our knowledge, the integration of the use of a general-purpose ASR into a CAPT tool, such as Google ASR, constitutes a novelty in the field of pronunciation training.

Initially, the first ASR launched by the company was the *Google Voice Search*[3] application in 2008. Exceptional improvements on the accuracy levels of previous speech recognition technologies were reported. Then, Google introduced elements of personalization into its voice search results, and used this data to develop its Hummingbird algorithm for the *Google Now* application in 2013. It arrived at a much more nuanced understanding of language in use. From 2016, the company released *Google Assistant*, an artificial intelligence-powered virtual assistant able to purchase products, send money, identify objects and songs, search the Internet, schedule events and alarms, among others. Although originally Google ASR was conceived as a smartphone application, nowadays it is available in other fields, such as driving systems, home virtual assistants and security systems. All Android devices can incorporate this ASR system for free since Google is the proprietary of this smartphone operating system.

The Google ASR system provides a hierarchically ordered n–best (n is defined by the user) list of probable sequence of words to match the input and a likelihood score (g–score) for each one. The main functionality is very simple: the user speaks and the system gives a string with the possible candidates in order, with the g–scores. Although its great performance and adaptability for developing purposes, the free version of the Google ASR system works as a black-box system. It does not allow to keep the recorded audio and the documentation is very limited.

However, at the end of summer of 2017 Google launched the beta version of a non-free **Google Cloud Speech-to-Text** service (GCSTT) and released its stable version (1.0) at the beginning of the 2018. This online product consists in a speech API which allows researchers to customize the ASR capabilities to a particular domain of the problem. Nine main features of GCSTT can be pointed out[2]:

[2] https://cloud.google.com/speech-to-text

[3] https://play.google.com/store/apps/details?id=com.google.android.googlequicksearchbox

1. **Automatic punctuation**: machine learning techniques grant to punctuate transcriptions accurately (i.e., periods, commas, and question marks).
2. **Global vocabulary**: a large words glossary of 120 languages and variants are supported.
3. **Inappropriate content filtering**: inappropriate text results can be filtered for some languages.
4. **Model selection**: four pre-built models are available: default, phone call, voice commands & search, and video transcription.
5. **Multichannel recognition**: audio recordings with two or more channels in which each speaker is in a channel (i.e., video conference or phone call) can be automatically separated and transcribed.
6. **Noise robustness**: audio recordings do not need to be pre-processed to handle noise problems.
7. **Phrase hints**: a set of words and phrases that are likely to be spoken can be given to the system to personalize and improve the results (i.e., custom words and names and voice-control use cases).
8. **Real-time streaming or prerecorded audio support**. Unlike the free ASR application, GCSTT allows users to keep the recorded file after recognition. It also provides a non-free platform for storing the audio files. Several audio encodings are supported, such as AMR, FLAC, and LINEAR16, among others. Speech recognition can be performed in three different ways:
   (a) **Synchronous recognition**: short (one minute length maximum) prerecorded audio samples can be processed in minimal time rates.
   (b) **Asynchronous recognition**: audio samples up to 180 minutes can be sent to be processed. Results can be periodically polled.
   (c) **Streaming recognition**: intended to process in real-time audio from a microphone, giving results while audio is being captured. It allows results to appear, for instance, while the user is still speaking.
9. **Speaker diarization**: automatic speaker identification is also possible.

Although a likelihood score (g–score) is also given with each candidate sequence of words in an ordered n–best list, the official documentation warns researchers that "This field is not guaranteed to be accurate and users should not rely on it to be always provided"[4]. Consequently, a rigorous process of adaptation of Google ASR and GCSTT service to the experiment prototypes has been required in order to maximize its scoring and diagnostic reliability (see Section 6.1.2).

Kaldi Speech Recognition System

In this thesis, the Kaldi framework is used for building a personalized ASR with different configurations and for analyzing comparative results with the general-purpose Google ASR system (see a guide to elaborate an ASR system with Kaldi in Appendix E). Six main features of Kaldi can be pointed out [167]:

[4] https://cloud.google.com/speech-to-text/docs/reference/rest/v1/speech/recognize

1. **Complete recipes**. They are intended to build speech recognition systems with broadly accessible databases, such as those supplied by the Linguistic Data Consortium (LDC)[5]: the Wall Street Journal Corpus, the Fisher-English Corpus, TIMIT, and more. Besides, they can serve as a template for training acoustic models on your own speech data.
2. **Extensible design**. Kaldi's algorithms are generic. The majority of functionalities are based on interfaces that allow to customize the code operations.
3. **Extensive linear algebra support**: Kaldi supports standard BLAS[6] and LAPACK[7] routines with a custom-built matrix library.
4. **Integration with FSTs**. The OpenFST toolkit is included as a library.
5. **Open license**. The code is available in GitHub and licensed under the permissive free software license Apache v2.0.
6. **Thorough testing**. Detailed and careful test routines are included in the majority of the source code.

6.4.2 Text-to-speech

Text-to-speech is a form of speech synthesis (artificial production of human speech) that converts text (input) into spoken voice (output) [168]. While *voice response systems* synthesize speech by concatenating sentences from a database of prerecorded words into fixed and invariable messages, TTS systems form sentences/phrases from scratch based on language's phonemes and graphemes [169]. In fact, TTS systems are theoretically capable of "reading" any string of word sequence to form original sentences.

As a general outline, three different stages can be distinguished in speech synthesis (see Figure 6.3). First, text-to-phoneme conversion, in which the text (rules, restrictions and dictionaries about sentences, words, phonemes, accents and stops, among others) is trained, analyzed, and processed. Second, the prosody modelling that includes intonation, rhythm, and intensity. A more natural and pleasant result for the user is achieved in this stage. Generally, this stage is diffusely shared between the text analysis module and signal generation one. Finally, the phoneme-to-speech conversion module, where the output signal speech is generated. It is based on the acoustic models and/or small units of pre-recorded wave-forms (i.e., concatenative synthesis, HMM-GMM based, DNN, hybrid approaches...). On the whole, the essence of TTS system design is to find a balance between flexibility, quality, and data.

[5] https://www.ldc.upenn.edu/
[6] http://www.netlib.org/blas/
[7] http://www.netlib.org/lapack/

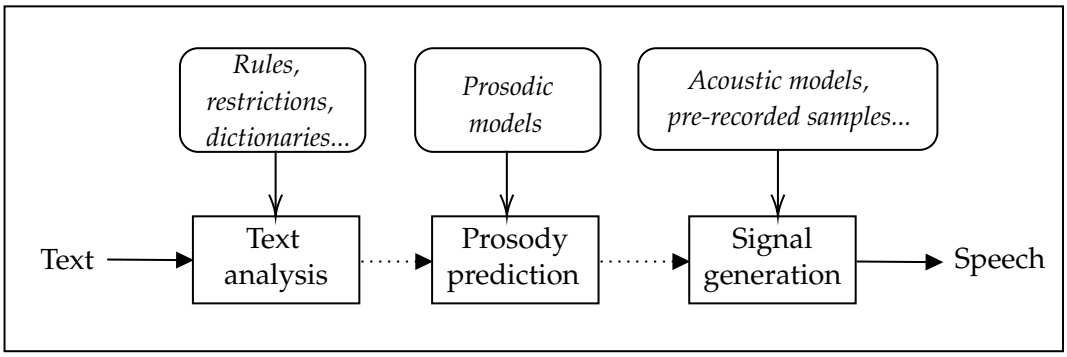

FIGURE 6.3: Generic architecture of a TTS system.

Giving more details about the design and building process of a TTS system exceeds the limits of this work. Interested readers can find an excellent revision of text-to-speech in [169] and details of speech prosody in speech synthesis in [170]. In this thesis, several features have been taken into account in order to select the TTS engine to be in our CAPT tools:

1. **Customization**: limitations about the number of words or languages and the possibility of changing speech characteristics, such as pitch and rate.
2. **Flexibility**: possibility of using vocabulary and phrases not employed for training during synthesis time. For instance, answering to unknown situations (i.e., a dialogue with the user or a very technical text). Also, it could refer to the adaptation to different voices or speech styles (i.e., reading a tale or a newspaper).
3. **Intelligibility**: quality of the audio generated.
4. **Naturalness**: human speech similarity. It is not required in all cases.
5. **Price**. Most of current off-the-shelf TTS systems are for free. Although it is possible to buy specific a whole TTS system, modern systems charge per transaction.
6. **Quality**: absence of noise and discontinuities, among others.
7. **Similarity to the original voice**. The TTS must capture the key features of human speakers successfully.

Google TTS technology

In the particular case of this thesis, the Google TTS[8] Android application has been employed for the experimentation. Seven main features of Google TTS motivate this election [171]:

1. **Audio format flexibility**. The audio can be generated in MP3, or LINEAR16, among others.
2. **Audio profiles**. The type of speaker from which the speech is intended to play can be selected (i.e., headphones or phone lines).
3. **Multilingual**. 180 voices and more than 30 languages and variants are supported.

[8] https://play.google.com/store/apps/details?id=com.google.android.tts

4. **Pitch and speaking rate tuning**. Up to 20 semitones more than the default output and up to 4x faster speaking rates are available.
5. **Text and Speech Synthesis Markup Language support**. Pronunciation instructions can be specified, such as numbers, pauses, and date and time formatting.
6. **Volume gain control**. The output volume can be adjusted from -96db to 16db.
7. **WaveNet voices**. They provide sounds "more natural than other TTS systems" [7].

6.4.3 Software Development

In this thesis, an incremental and iterative development of the engineering methodology has been followed [172]. This methodology consists in developing initial versions of the CAPT tools, also called **prototypes**, in which successive improvements are applied, improving their quality of until the final version. The main phases in this methodology are planning, analysis and design, implementation, testing, deployment, and evaluation. Furthermore, a modular software design has also been implemented following a version control system (Git), in which the software code has been reused in the next versions of the CAPT tools (see Table B.1 in Appendix B for an estimation of the number of software development days in this thesis).

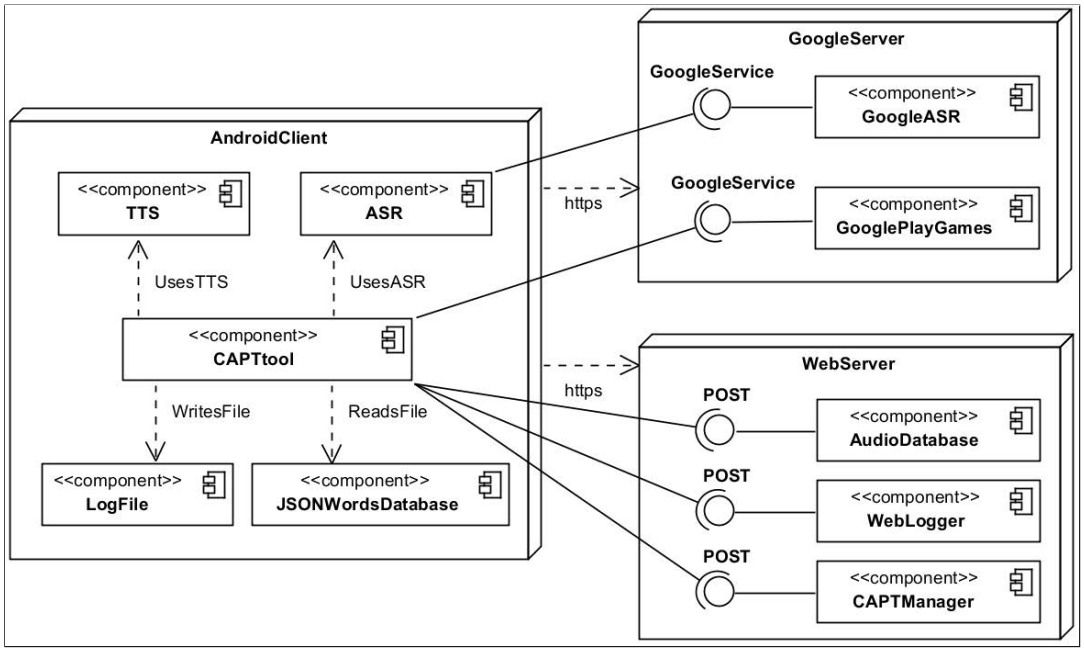

FIGURE 6.4: Client–server model of the prototypes of this thesis, adapted from [31].

All the CAPT tools developed in this thesis have in common a client–server architecture (see Figure 6.4). The client is an Android device (version 4.4 or higher) in which the CAPT tool is installed; whereas the server part is divided into two components: external (Google) and personalized services (private own web server). The interaction results between the user and the CAPT tool in the client are saved as a JSON format in log files that compile all possible depersonalized data diachronically. They are sent to our web logger in the server part automatically. These log files contain the same structure of data fields for each experiment, including new fields when

necessary. The audio files are also sent to the web server. The lists of minimal pairs elaborated by experts are defined in a file (*JSONWordsDababase*) which includes in each line the orthographic transcription, phonetic transcription, and the possible homophone words of the minimal pairs. The Android client must have access to the ASR and TTS technology integrated in the CAPT tool. Finally, the game versions of the tool use the Google Play Games platform which provides gaming service and software development kits of ready-to-use game features in the software applications. This platform is complemented by our *CAPTManager* component, specifically developed for our CAPT tools. Some of its functionalities are the user's log-in and sign-up, and the scoring system, among others.

For each one of the experiments, a CAPT tool for smart devices has been developed (see the download links in Section 9.3.7). These tools can be also run in desktop PCs by means of Android emulators. The iOS operating system was out of scope of this thesis. In particular, the CAPT tools have been developed with Android Studio version 3.0, the software development kit for Android version 26, and Java version 6. Six academic projects of the University of Valladolid have been related directly [173], [174], [175] or indirectly [176], [177], [178] with the experimental design of the tools, and other three more are currently being carried out. The particular characteristics of the technology employed during the experimentation taken into account have been (see a comparison in Table B.4 in Appendix B):

1. **Smart devices** that support Android version 4.4. or higher with full access to the Internet and a minimum system's storage of 500 megabytes. The prototypes are installed in this instrument.
2. **Device's OS**: Android version 4.4. or higher; or Windows 7 or higher (with NOX App Player 5.0.0 or higher[9] support).
3. **Data server system**: Linux standard distributions 2.6 or higher, derived from GNU/Linux; Windows Server platforms 2008 r2 or higher; or Mac OS X 10.6 or higher. Minimum system's storage of 10 gigabytes. The server aims at gathering all statistical data of user's interaction with the system, to provide some services through the Internet, and to keep audio files when needed.

6.5 Corrective Feedback Mechanisms

In this section the main relation between the CF mechanisms included in the prototypes of the experiments presented in this thesis and those in the literature is pointed out (see a comparison in Table B.8 in Appendix B). The different CF strategies followed in each experiment of this thesis are enumerated at the end of this section (the definition of each one of them is available at Chapter 4). First, in all the experiments carried out in this thesis written visual feedback is provided to users, not only via the orthographic representation (like the majority of studies reported in Section 4.2) but also via the phonetic transcription of the words following the International Phonetic Alphabet (IPA) [179].

Explicit feedback is given to the learners in the discrimination exercises likewise [13], [37], [82], [85]. That is, the chosen word is highlighted in green color with a sound and a message of success. When the answer is wrong, it is highlighted in

[9] https://en.bignox.com/

red with a sound and a message of failure. Along the experiments presented in this thesis, the word to choose is always synthesized by a TTS.

The possibility of listening to models of words as an optional explicit feedback, but using a synthetic voice [97] instead of a natural voice as in [13], [82], [87], [88], [89], [90] have been offered to the learners in all the experiments. As a novel contribution, if the learner does not reach the minimum score expected, the CF suggests a set of word exposures. It consists in listening to three minimal pairs of the same phoneme of the word up to five times, in normal and reduced speed, alternatively. Then, the individual returns to the previous failed training mode.

Regarding production activities, several feedback techniques have been applied. The ASR-based CAPT system provides assessment of the utterance spoken. Besides, it responds to the user's pronunciation of each intended word with a right/wrong sound and by changing its color to green or red. In addition, after a wrong attempt, instead of highlighting the mispronounced part of the word like in [13], a message containing the sequence of most probable words (n–best list) recognized by the ASR and the number of remaining attempts is displayed, requesting a new utterance. In order to avoid students' frustration, the number of attempts per word is limited [13], [65]. Error-based feedback that goes beyond the mere iteration of trial–and–error cycles is also implemented [13], [64] and simple right/wrong or good/fair/bad feedback [60], [62], [65]. As a novelty, the system executes an explicit corrective feedback response that invites users to listen to the synthesized version of the problematic word. Given a determined number of consecutive failures after finishing all training activities, as implicit feedback, the number of correct and wrong answers are shown with a final score, a smiley/sad-face emoticon, and a chime/buzzer sound, respectively [13].

Furthermore, in some prototypes carried out in this thesis, the first way of explicit feedback provided is a brief video with theoretical concepts and practical examples, presented in each pair of phonemes to be contrasted, similar to [13], [63], [64]. In particular, each video consists in an exposure to the minimal pairs of the contrast with articulatory descriptions, perception cues, and instructions of how to turn L1 sounds into an L2 sounds. Sagittal planes and animation mouth shapes are also included (see more details about this particular videos in Section 6.2.1). Besides, the idea of providing advises or tips for helping learners after wrong production attempts [61] is being integrated in the latest prototypes of this thesis for future work (see more details in the last paragraph of Section 6.2.1).

To sum up, the different mechanisms of CF applied in each prototype of the experimentation of this thesis are:

- **Implicit CF**: repetition's request of a mispronounced utterance with neither the recognized words nor hints; word's phonetic transcription; right/wrong answer sounds; interface color changes; activity score; happy/sad smiley after an activity's performance; and next exercise recommendation.
- **Explicit CF**: repetition's request of a mispronounced utterance with the recognized words and a hint; explicit correction of the activity; word synthesis; dual listening to synthesized and own utterances; and theoretical-practical video with sagittal planes and animation mouth shapes.

6.6 Game Instruments

In this section the game instruments and strategies included into the CAPT systems developed for the prototypes of the experiments are presented in order to promote motivation and competitiveness to learners (see Chapter 5 for specific details of each one of them, and Tables B.5 and B.6 in Appendix B for a comparative). These game elements can be categorized as follows:

- **Approach**: individualistic or social (implicit or explicit competition) activities.
- **Points**: a score assigned to each activity. It depends on the difficulty and number of attempts performed per task.
- **Leaderboards**. The points obtained for performing activities contribute to climbing up a ranking and acquiring specific experience.
- **Badges**: achievements (digital trophies) and motivational messages categorized by languages and type of activities.
- **Prizes**: an extrinsic motivation for achieving a final goal (i.e., a reward or a diploma).
- **Performance graphs**: individual or social information about user's results along the experiment.
- **Avatars**: representation of the user profile in the tool.
- **Restrictions**: timers, attempts per activities, choosing the wrong word in perception exercises, specific matchmaking, and clear tickets.
- **Progress**. Each task result is showed as the final result of an activity or lesson, unlocking new content.

While in some experiments the Google Play Games platform[10] has been used to include some gamification elements, in others an own platform from scratch has been developed.

6.7 Selection of Participants

The main target population of this thesis were university students from the Language Learning Center of the University of Valladolid who voluntarily accepted to participate in the experimentation. They were invited to take part of the experiment via invitation emails to their corporate university email address. Students who agreed filled in a registration form with their demographic information and signed an informed consent in which the data gathering, location, and schedule of the experiment were detailed. Individuals were rewarded with a diploma, an academic certification, or a prize.

Participants were characterized and grouped by their homogeneous L2 ability. The number of women and men was balanced when possible. Subjects were classified by their L2 proficiency and in different groups, depending on the experiment. Interested readers can find specific details of demographic and characterizing information in Table B.7 of Appendix B and in the corresponding section for participants description of each experiment.

[10] https://play.google.com/store/apps/details?id=com.google.android.play.games&hl=en

6.8 Summary

There are some pedagogical and technological decisions to take into account for elaborating a CAPT system for a particular L1 group of learners, such as its methodological activities (free selection or guided), the pronunciation improvement assessment, the speech technology employed, the feedback given to the users, and the inclusion of gamification elements.

One of the main contributions of this dissertation is the specification of a innovative protocol for selecting adequate minimal pairs lists for personalized CAPT systems. Another one is the specific flow of personalized activities presented to the user (a cycle of exposure–perception–production activities), partially based on the NCM. Thirdly, the possibility of integrating off-the-shelf ASR and TTS systems in a non-obstructive way is also a novel fact in the CAPT literature. Finally, the inclusion of gamified elements in a competition about pronunciation is also a novelty in the field of L2 pronunciation training.

In this chapter the common elements of the dimensions of the experimentation has been introduced. First, the importance, characteristics and limitations of minimal pairs to take into account for L2 pronunciation training have been described. Then, a novel protocol for selecting minimal pairs lists for specific L1 and L2 and integrating them with speech technology in a CAPT system has been detailed. Second, each one of the training activities adopted for the prototypes have been detailed. That is, the adaptation of the NCM and other related pronunciation training programs and the fundamentals for selecting the specific activities. Third, the objective and subjective strategies adopted to assess user's pronunciation improvement have been described. Fourthly, the fundamental principles of open-source, semi and full commercial off-the-shelf ASR systems have been detailed, such as their desirable characteristics, their mathematical fundamentals, their architecture, and their performance metrics. Then, two state-of-the-art ASR technologies (Google's speech recognition and Kaldi), including a general outline for building a personalized ASR system from scratch have been analyzed. It has also been presented an overview about text-to-speech technology, its current applications, its preferable characteristics and examples of state-of-the-art TTS systems. Fifth, the corrective feedback mechanisms adopted to the CAPT systems have been explained Sixth, the main gamification strategies followed in the experimentation have been described. Finally, the common characteristics of the experimentation participants have been mentioned.

Chapter 7

Experiments

This chapter focuses on the presentation of the experiments carried out in this thesis to answer the research questions and to validate the research objectives established. Four different experiments with real users and scenarios within different academic institutions were accomplished along an evolutionary process to incorporate different alternatives. Current ASR and TTS systems are integrated in all of them. The training activities methodology for improving L2 pronunciation is refined along the evolution of the four tools, including an individualistic and a social training approach. The feasibility of a guided and non-guided training protocol is also explored. Different feedback strategies are tested and improved with each experimental iteration. Besides, different gamification elements are incorporated to evaluate their influence on the training procedure. As part of the experiments, experimental data is automatically gathered for future analysis, including the interaction results with the CAPT system in all experiments and the audio files obtained during the experimentation when possible.

The first section of this chapter presents the main reasons that motivate the strategy followed for the experimentation, including an overview about the main characteristics of each case of study. The rest of the sections in this chapter describe the specific details of the design and methodology followed in each experiment, and the most relevant results achieved are reported to validate the research objectives of the thesis.

7.1 Experimentation Roadmap

An incremental and evolutionary approach has been followed in the design of the four experiments presented in this thesis, so that the results of each of them have been analyzed to change and refine the design of the next experiment. This work has been part of three different research projects developed by multidisciplinary teams and their objectives have also influenced this thesis' steps and experimentation flow (see Section 9.3.9 for more details about funding).

Five prototypes have been developed for the four experiments of this dissertation as shown in Figure 7.1. The evolution of these prototypes aligns with two main focuses: the first focus (the three experiments at the bottom of the figure) proposes incorporating gamification elements and social strategies; whereas the second focus (the three prototypes of the experiment at the top of the figure) follows an individualistic approach with guided training instructions.

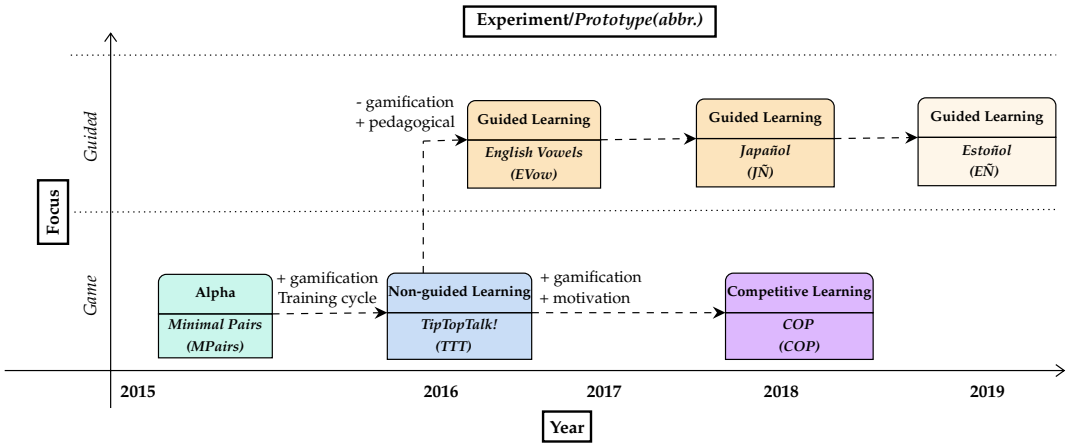

FIGURE 7.1: Evolution diagram of the experiments and prototypes of this thesis.

The first experiment of this dissertation (***Alpha***) targets the evaluation of the inclusion of state-of-the-art speech technology (ASR and TTS) into an activity protocol for pronunciation training. In order to give answer to the research question RQ1 (and *Issue 1.1*, see Section 1.3 for more details about the research questions and their issues), a set of steps are identified (RO1, RO2, and RO3). In particular, and as a novelty, two general-purpose speech technology systems (Google's TTS and ASR systems) are tested by different groups of real subjects with isolated production exercises of minimal pairs words which are selected by linguistic experts. A mobile CAPT system is designed and developed from scratch in which all user's interaction data is gathered automatically. Three different groups of users according to their *a priori* English pronunciation level (according to the Common European Framework of Reference for Languages, CEFR[1]) participate in this experiment (see Section 7.2 for more details).

The second experiment (***Non-guided Learning***) aims at assessing the possible learners' pronunciation improvement along time within an individualistic competition, using a gamified CAPT system with ASR and TTS technologies, named TipTopTalk!. In addition to the research question RQ1 (and *Issue 1.1*), it is also intended to give answer to RQ2 (and its *Issues*) and RQ3 by reaching the research objectives RO1, RO2, RO3, and RO4. It follows an innovative approach in terms of the recommended training cycle of activities and the gamification elements included to motivate and encourage learners to keep on training. American English and Spanish native speakers with different pronunciation levels joined it (see Section 7.3 for more details).

A production improvement stagnation was detected in the most proficient learners of the second experiment. The students were not provided with enough resources to solve their pronunciation mistakes (i.e., the isolated use of TTS technology), so that they lost motivation, and tended to train the easier exercises to obtain positive outcomes. These reasons motivated a new focus of research, more pedagogical, guided, and individualized. The third experiment of this thesis (***Guided Learning***), tries to give answer to the research questions RQ2 (and its *Issues*), RQ1, and *Issue 1.1*, taken into account the research objectives RO1, RO2, RO3, and RO4. The main objective is to train user's pronunciation by guiding her/him through a

[1] https://www.cambridgeenglish.org/exams-and-tests/cefr/

CAPT system with a personalized and a more precise feedback, based on learner's results (see Section 7.4 for more details).

Two prototypes are developed for the third experiment, named ***English Vowels*** and ***Japañol***. Castilian Spanish and American English are the L1 and L2 target languages for the former one. Participants belong to intermediate EFL courses of the Language Center of the University of Valladolid. Japanese and Castilian Spanish are the L1 and L2 target ones, respectively, of the Japañol prototype. Participants belong to intermediate L2–Spanish courses for Japanese students of the Language Center of the University of Valladolid and University of Seisen. As a consequence of the collaboration with the University of Tartu, Estonia, at the end of the thesis a new experimentation phase started (*Estoñol* prototype). It follows the same philosophy of English Vowels and Japañol prototypes. In particular, its L1 and L2 targets are Spanish and Estonian, respectively [30]. Native Spanish learners of Estonian from University of Tartu participate in this prototype. A pre/post test strategy to ascertain the pronunciation level improvement of the participants from different training groups (experimental, in-classroom, and placebo) is followed, in a similar way to the other prototypes of this experiment.

In the light of the results of the second experiment, and almost in parallel with it, a competitive approach is carried out in the fourth experiment (***Competitive Learning***). All research questions defined for this thesis are tried to be answered taking into consideration the research objectives RO1, RO2, RO3, and RO4. In particular, a second version of the app developed for the second experiment (TipTopTalk!) is developed, in which learners can challenge each other under a set of common rules —unlike the competition of the second experiment, in which learners cannot challenge other users (single-player, individualistic approach). We try to overcome the pronunciation stagnation detected on the most proficient players. The competition-based configuration allows to gather a great number of utterances and user's behavior data with this gamified CAPT system. University students with different proficiency level of English as L2 take part in this experiment (see Section 7.5 for more details).

Table 7.1 shows the main characteristics of each experiment in terms of methodology, technology, game elements, and kind of assessment. In some cases these features are shared among the experiments, but in others, they are unique. The main common elements included in all experiments are the minimal pairs lists included in each training activity, log files for gathering all training data, an ASR system for (1) converting learner's utterances audio to text and (2) the assessment of production training activities, a TTS system for synthesizing words, and a set of training activities for assessing/improving pronunciation. Finally, a set of exposure, discrimination, and mixed activities are defined from the TipTopTalk! prototype.

On the other hand, since the TipTopTalk! and COP prototypes are based on competitions, they include some gamification elements, such as a leaderboard, a competition strategy, and trophies. The qualitative research techniques, focus group, and questionnaires, are also carried out in these two prototypes. In particular, several questionnaires about motivation, pronunciation level, attitude toward competition, and reasons for abandoning are included in the COP prototype. This prototype also gathers user's utterances in the training activities and in the pre/post-quests. It also has the peculiarity of allowing users to challenge each other with a limited quantity of pronunciation activities.

	Exp. 1	Exp. 2	Exp. 3		Exp. 4
	Minimal Pairs	**TipTopTalk!**	**English Vowels**	**Japañol**	**COP**
Methodology					
Minimal pairs	X	X	X	X	X
Production activities	X	X	X	X	X
Training activities	X	X	X	X	X
Discrimination activities		X	X	X	X
Exposure activities		X	X	X	X
Mixed activities		X	X	X	X
Guided protocol			X	X	
Theoretical-practical video			X	X	
Technology					
Log files	X	X	X	X	X
ASR	X	X	X	X	X
TTS	X	X	X	X	X
Audio recordings				X	X
Game instruments					
Leaderboard (game points)		X			X
Competition		X			X
Badges (trophies)		X			X
Challenges					X
Limited activity selection					X
Assessment					
Focus group	X				X
Pre/Post-tests			X	X	
Questionnaires		X			X

TABLE 7.1: Main elements included in each prototype of the experimentation classified by categories.

Finally, a guided training protocol is the pedagogical approach followed in the English Vowels and Japañol prototypes. Audiovisual materials with theoretical explanations about the nature of the phonemes within each minimal pair and practical illustrations of the sounds of these phonemes are included, and a pre/post-test strategy for assessing user's pronunciation improvement before and after training is followed.

7.2 Alpha Experiment

Alpha is the name of the first experiment carried out, and a mobile learning application was developed from scratch as a means to provide L2–English pronunciation assessment for native Spanish speakers, named *Minimal Pairs*. It contained a minimal pairs set selected by a phonetics expert which users had to confront by producing their word utterances with general-purpose speech technology integrated into the learning application. This experiment was a starting point for discovering the weaknesses and limitations of current ASR and TTS technologies. Categorizing speakers by assessing their English pronunciation level with the help of ASR and TTS technology, from basic to native level, was also possible.

7.2.1 Experimental Procedure

The recruitment campaign lasted for 7 days. Learners who agreed to participate followed a one-session training protocol, as shown in Figure 7.2. All subjects (see Section 7.2.3 for more details about participants) were asked to perform the same pronunciation training activities with the system during a maximum time of 7 minutes, individually. The training session was carried out in a quiet testing room that contained a comfortable chair and a small table with a tablet in which the CAPT system software was installed. Before starting the session, a member of the research team gave instructions of use to participants. Then, each speaker had to perform the activities proposed by the system. All the interaction events, timestamped, and the results of the ASR engine for each attempted word were stored into log files for later analysis. Furthermore, a one-hour focus group session with some randomly-selected non-native participants of the experiment was conducted.

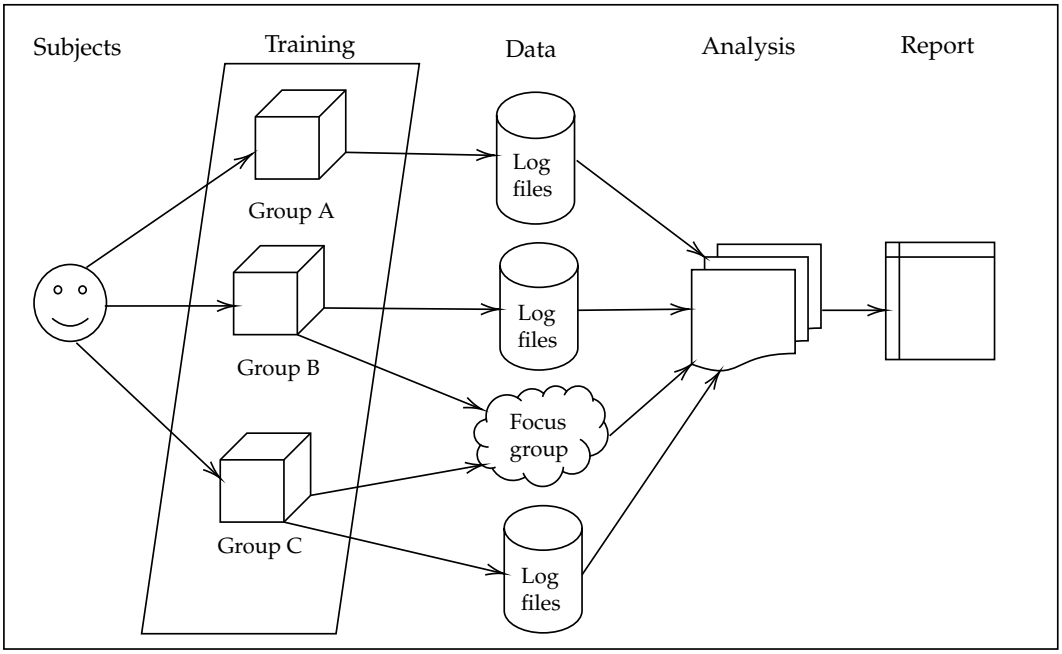

FIGURE 7.2: Steps of the first experiment's protocol.

7.2.2 Enrollment

There were three different recruitment campaigns for this experiment. First, a group of American native learners of Spanish of the same course at the Language Center of Valladolid were asked to participate voluntarily by attending to their classroom. Their aim was to test the feasibility of the system. Second, students from the English philology degree of the University of Valladolid were invited to take part in the experiment via email. Finally, students from the Computer Engineering degree of the University of Valladolid were also asked to participate via invitation emails. These two groups of Spanish students were the main subjects for this study, while the American students were incorporated to test speech technology adequacy.

Students filled in an agreement and a registration form with their demographic information. A specific time slot was reserved for each participant to perform the training activities individually. All participants were awarded with a diploma after

completing the protocol, and a reward was given to those who also attended the focus group.

7.2.3 Participants

Users were divided into three different groups, according to their English (en_US) pronunciation proficiency level:

1. **Group A**: 12 native American speakers between 18 and 26 years old. 5 were women and 7 were men. They all belong to the same L2 Spanish course at the Language Center of the University of Valladolid.
2. **Group B**: 21 undergraduate students of English Philology at the University of Valladolid between 18 and 26 years old. 11 were women and 10 were men. They all claim a C1–C2 English proficiency level as L2 and have passed the same advance English phonetic course.
3. **Group C**: 20 Computer Engineering students from the University of Valladolid between 18 and 26 years old. 6 were women and 14 were men. Their English proficiency level as L2 was lower than the rest of participants (B1–B2).

All participants took part in the same testing activities of the experiment. It was expected group A achieved the best results, and group C the worst ones. Some randomly selected speakers of Group B and Group C took part in the focus group session.

7.2.4 Minimal Pairs CAPT System Description

The training activities were performed by the speakers using a CAPT system developed from scratch, called *Minimal Pairs*. It is a software tool for smart devices which presents twelve American English minimal pairs of vowel and consonant contrasts. They are randomly chosen from a set of twenty difficult pairs for Spanish speakers selected by a phonetics expert.

Participants had to produce the words correctly, so that the words were recognized by Google ASR. A production attempt was considered correct (right) when the orthographic transcription of the word (or some homophone) was included in one of the first five positions of the text hypotheses of the ASR result. Five attempts maximum were allowed. Participants could also listen to the synthesized form of the words (Google TTS) as feedback. Data related to user-system interaction was gathered via log files.

Figure 7.3 shows a screenshot of the Graphical User Interface (GUI) of the CAPT system. For each minimal pair of the training session, both words are shown close to a representative image of each one. There is a button on the right side of each picture to listen to the synthesized word. The remaining time, the number of attempts per word, the number of remaining pairs, and the number of correct/wrong utterances are also displayed. There are three buttons at the top of the figure to go back to the previous pair, to go forward to the next one and to finish the training session. Instructions are written at the bottom of the screen.

Figure 7.4 shows the result of a user interacting with the system with a minimal pair (the final state of the Figure 7.3). The speaker has correctly uttered the first word of the minimal pair and the interface has changed its main color to green, disabling the possibility of producing again the word. On the other hand, the second word

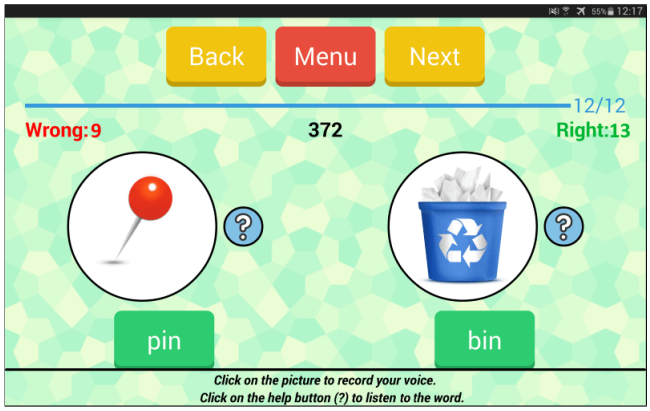

FIGURE 7.3: Production activity GUI of the Minimal Pairs prototype (before).

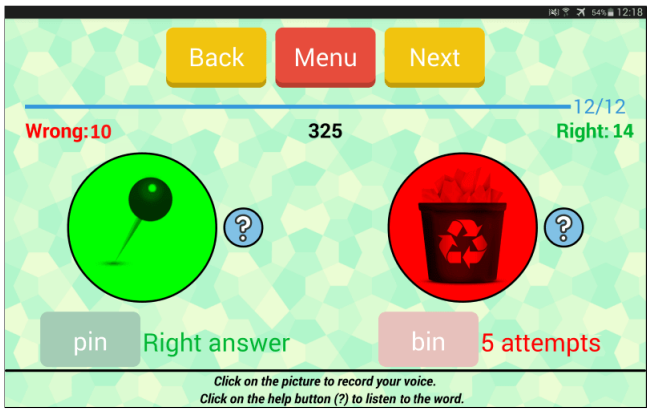

FIGURE 7.4: Production activity GUI of the Minimal Pairs prototype (after).

has been disabled because the user has not been able to produce it correctly in five attempts. In this case, the main color of the word interface is changed to red.

7.2.5 Instruments and Metrics

There were three different sources of data:

- **Registration forms**: user's demographic information, such as name, age, gender, L1, academic level, and final consent to analyze all gathered data. This information was carefully collected and saved into physical text documents.
- **User's interaction log files**. The CAPT tool gathered data associated with all low-level interaction events and monitored all user activities. This data was saved into local log files and automatically uploaded to a web server. From these, a set of experimental variables were identified and computed:
    1. **Training intensity**. Which computed the amount of events tracked in the experiment. It was derived from the number of discrimination and production tasks; the number of times a particular phoneme was practiced; and the times a word was listened to.

2. **Training performance**. Which measured the number of successful answers obtained by the participant for each tracked activity during a specific time. The variable encompassed right and wrong discrimination tasks; right and wrong pronunciation tasks (with the n–best list of hypotheses and g–score values, see Section 6.2.4); success rates in discrimination and production tasks per phoneme; and time spent on performing training events.

- **Focus group session**: The audio of the session was recorded via a camera and the most important opinions and requests of the participants are written by a member of the research team by taking notes. This meeting was carried out in a classroom with all participants face to face. One member of the research group conducted the session while other took notes and recorded the audio of the meeting. Firstly, an overview of the results obtained during the experiment was presented to the participants during 15 minutes. Then, subjects were asked about their perceptions and opinions about the experiment with the possibility of discussion with other participants (30 minutes). The last 15 minutes of the session were intended to suggest improvements and future work.

7.2.6 Results

The first research question of this thesis about the inclusion of ASR and TTS systems in a CAPT tool (RQ1 and *Issue 1.1*) was tried to be answered with the results obtained in this experiment following the steps defined by the research objectives RO1, RO2, and RO3. These results are presented in next paragraphs according to their origin. That is, (1) results from the interaction between the students and the CAPT system during the training session and (2) results from the focus group session. The discussion of these results is included in Chapter 8. Results related to the training session have been partially published in [17].

User's Performance

Table 7.2 shows the total number of production attempts with the ASR system ($\#ASREvents$), the total number of listenings with the TTS system ($\#TTSEvents$) and the total time spent with the system ($Time(s)$). In terms of time to complete activities, group C was the slowest since its members used the TTS to listen to a word (requested listening) more times (606), and also, the number of attempts to produce a correct word with the ASR was higher (1094). Globally, the ASR system was used 2.42 times more than the TTS one, even though the TTS events were not restricted to a particular maximum number of events per word (as a reminder, students can produce a word with the ASR system up to five times). Statistically significant differences were found between the three groups and the three variables $\#ASREvents$, $\#TTSEvents$, and $Time(s)$ of Table 7.2, as determined by one-way ANOVA test [180] ($p < 0.001$, 95% confidence level in the three cases). In particular, pairwise group comparisons were carried out to examine these differences in pairs. A t-test [181] at 95% confidence confirmed that there were statistically significant differences between all group pairs ($p < 0.001$) except for the $\#ASREvents$ column regarding Group B and Group C ($p = 0.06$).

Regarding the results related to the ASR system, Table 7.3 shows the mean number of production events with the ASR system ($\overline{\#ASREvents}$ column), the mean number of correct productions ($\overline{SuccessASR}$ column), the mean number of wrong

Group	# Speakers	# ASREvents	# TTSEvents	Time (s)
A	12	372	35	2431
B	21	1033	400	6677
C	20	1094	606	7492
Total	53	2499	1041	16600

TABLE 7.2: Descriptive data gathered with the Minimal Pairs CAPT system, adapted from [17]

productions ($\overline{FailASR}$ column) and the mean percentage of success comparing the number of times the ASR system identifies as correct a word with the number of production attempts of such word ($\overline{Recall}$ column). Group A achieved the best results since its participants take -on average- less production attempts (31±7 of a total of 120 maximum attempts; five maximum attempts for each one of the 20 words of the 12 minimal pairs presented in the activity). Besides, Group A reached the highest $\overline{SuccessASR}$ rate, the lowest $\overline{FailASR}$ one and the best $\overline{Recall}$ rate. Thus, the higher the declared L2 level, the better production results with and without repetition. On the other hand, Group C reached a 73% of wrong attempts with the ASR and achieved the worst results in all rates.

Statistically significant differences were found between the three groups and the four variables of Table 7.3, as determined by one-way ANOVA test ($p < 0.001$, 95% confidence level in the four cases). Pairwise comparisons with t-tests at 95% confidence were also run for all variables of the table, confirming that there were differences between all of them except for the $\overline{\#ASREvents}$ column regarding Group B and Group C ($p = 0.06$). The differences between the $\overline{SuccessASR}$ rate of the participants of Group B and the rest of participants deserve attention. An ANOVA test and a t-test at 99% confidence confirmed that there were statistically significant differences in the same cases as 95%. except for the $\overline{SuccessASR}$ rate between Group A and Group B ($p = 0.057$).

Group	$\overline{\#ASREvents}$	$\overline{SuccessASR}$	$\overline{FailASR}$	$\overline{Recall}$(%)
A	31±7	21±4	10±6	69.1±17
B	49±14	18±3	31±15	41.2±15
C	55±9	15±4	40±10	28.1±10

TABLE 7.3: ASR-related results gathered with the Minimal Pairs CAPT system, adapted from [17]. Values after the symbol ± represent the standard deviation.

As explained in Section 6.2.4, an n–best list of predictions was provided by the off-the-shelf Google ASR system in each utterance (in this experiment a 5–best list). This list consists of pairs of a text and a numerical score, also called g–score, with values in a scale of [0, 1]. The next two tables report the results related to this issue.

First, in Table 7.4 the mean g–score value of the right production attempts ($\overline{Right}$ column), the mean g–score value of the wrong production attempts ($\overline{Wrong}$ column), the mean g–score value of any production attempt ($\overline{Total}$ column) and the mean time spent with the tool each user ($\overline{Time}$ column) were represented, categorized by groups. Statistically significant differences were found between the three groups

and the four variables of Table 7.4 ($p < 0.001$, 95% confidence level in the four cases). In particular, pairwise comparisons with t-tests found differences in all cases ($p < 0.001$, t-test, 95% confidence), except for Group A and Group B in the case of wrong attempts ($p = 0.09$, t-test, 95% confidence). That means native speakers -on average- need less time to perform the activities than advanced learners and beginners, respectively, as intuited in Table 7.2. Group C speakers were the slowest ones since they need a high number of production attempts due to their wrong utterances (see Table 7.3) and the quality of the ASR response was not as much confidence as the other groups (0.55 vs. 0.59 vs 0.59). Besides, their utterances achieved better g–score values in the majority of cases, except when comparing wrong attempts values in Group A and Group B, which was similar without significant differences, as mentioned earlier.

	g–score			
Group	$\overline{Right}$	$\overline{Wrong}$	$\overline{Total}$	$\overline{Time}$ (s)
A	0.70±0.3	0.59±0.3	0.67±0.3	203±66
B	0.65±0.3	0.59±0.3	0.61±0.3	318±82
C	0.58±0.3	0.55±0.3	0.56±0.3	375±54

TABLE 7.4: ASR-related metrics gathered with the Minimal Pairs CAPT system, adapted from [17]. Values after the symbol represent the standard deviation.

Second, Table 7.5 shows the distribution of the target word of an utterance when it was included in the ASR 5–list of results. The Group A production quality was higher than the rest of Groups since the target word was recognized in the first position of the results in 63.6% of the attempts, in contrast to Group B (51.8%) and Group C (47.3%). A Chi–square [182] test confirms the statistically significant differences between the three groups regarding the first position of the results ($p = 0.0053$, $\chi^2 = 21.7869$, df = 8, at 95% level).

	Position (%)				
Group	1st	2nd	3rd	4th	5th
A	63.6	18.8	8.4	6.8	2.4
B	51.8	21.8	13.7	10.6	2.1
C	47.3	23.8	13.8	9.7	5.4

TABLE 7.5: Mean distribution of the target word in each recognized utterance with the Minimal Pairs CAPT system, adapted from [17].

Table 7.6 sheds light about the reasons why native speakers also fail with the CAPT system when producing words in their L1, as shown in previous Tables. The list of 20 minimal pairs for this first experiment was designed by a phonetics expert in American English, based on his experience in the field after his teaching years. This list of words was directly integrated into the CAPT system. Results showed problems with infrequent in everyday English words, such as *wreathe*, *luff*, or *wader* which summed the 50% of the total wrong utterances of the experiment. Besides, the word *wreathe* was never identified by the CAPT system and the word *luff* was reported correct in only two events. The fifteen most frequently failed words in Groups B and C account for the 70% of the attempts. However, in the case of native

speakers (Group A), this value was not reached after the word *wader*. Furthermore, this table shows words very frequently confused by Spanish speakers, such as *peck*, *Dawn*, or *sue* that were never confused by native speakers. In Chapter 8 we will discuss the consequences of and actions to take to improve the selection of words and speech technology proper for a CAPT system.

	Group A		Group B		Group C	
	Word	%	*Word*	%	*Word*	%
1	wreathe	100	luff	100	wreathe	100
2	luff	94	wreathe	100	luff	98
3	wader	73	letch	97	letch	98
4	soot	64	loose	90	wader	96
5	sock	58	wader	88	sock	96
6	caber	56	peck	84	soot	96
7	letch	50	sue	84	Gwen	89
8	mass	38	sock	83	shun	88
9	don	33	dunce	81	sue	86
10	mess	33	dawn	80	dawn	85
11	Gwen	31	soot	79	were	83
12	shun	30	Gwen	76	peg	83
13	were	20	were	72	peck	82
14	dunce	12	don	71	loose	81
15	mat	11	zoo	70	dunce	81

TABLE 7.6: Most frequently unrecognized words by the ASR system in the Minimal Pairs prototype (in percentage), adapted from [17].

Finally, regarding the results related to the use of the TTS system as feedback for the production of words, Table 7.7 shows the average number of listenings to each word of the experiment with the TTS system ($\overline{\#TTSEvents}$ column), the mean percentage of correct productions after listening ($\overline{SuccessTTS}$ column), the mean number of wrong productions after listening ($\overline{FailTTS}$ column), and the mean percentage of the number of times a learner used the TTS system with respect to the total number of listening and productions events ($\overline{Rate}$ column). Users listened to the synthesized models of the words when they had doubts about the way to produce them. The number of times they could synthesize the words was not limited. In particular, in Table 7.7 the results related to the words *wreathe* and *luff* were not included since these words were found the most problematic ones with the ASR system in this experiment as explained in Table 7.6.

The use of the TTS by natives was negligible (2±2 on average). They only resorted to it when the system did not identify their utterance, being only a 5.5% of the production and synthesis events. Besides, their $\overline{SuccessTTS}$ rate was the worst and their $\overline{FailTTS}$ rate was the highest one, in comparison to the rest of participants. That means TTS feedback was not helping natives with the not-recognized words by the ASR. Although non-native speakers used the TTS more times than natives (27.6% and a 35.5%), the feedback provided seemed to be not sufficient enough since non-natives' $\overline{SuccessTTS}$ rate values were low and similar to natives ones (30.3% vs.

29.4% vs 26.3%). In particular, Group C participants made use of the TTS system more than the rest of the groups (28.0% vs. 18.0% vs. 2.1%, respectively), including production and synthesis events (35.5% vs. 27.6% vs. 5.5%, respectively). Statistically significant differences were found between the three groups and the four variables $\overline{TTSEvents}$, $\overline{SuccessTTS}$, $\overline{FailTTS}$, and $\overline{Rate}$ of Table 7.7, as determined by one-way ANOVA test ($p < 0.001$, 95% confidence level in the three cases). Pairwise comparisons with t-tests confirmed these results in all cases ($p < 0.05$) except for the $\overline{FailTTS}$ and the $\overline{SuccessTTS}$ rates between Group B and Group C ($p = 0.08$, at 95% confidence).

Group	$\overline{\#TTSEvents}$	$\overline{SuccessTTS}$(%)	$\overline{FailTTS}$(%)	$\overline{Rate}$(%)
A	2.1±2	26.3	73.7	5.5
B	18.0±12	29.4	70.6	27.6
C	28.0±17	30.3	69.7	35.5

TABLE 7.7: TTS-related results gathered with the Minimal Pairs CAPT System, adapted from [17].

Focus Group Session

This meeting was carried out with 10 non-native learners who participated in the test session. The notes extracted from the participants' impressions, opinions and improvements about the ASR system were mainly positive. They supported the results presented in the previous Tables 7.2, 7.3, 7.4, 7.5, and 7.6, in which non-native learners achieved the worst results in production activities. However, some comments, such as "*I would like to hear and compare my utterances with the ones in TTS*", "*I felt frustrated after continuous failures*", "*I would like more training activities. For example, be able to listen to a word of a minimal pair and select which one is the correct answer*", strengthened the idea of the necessity of designing and including new non-isolated training activities and corrective feedback techniques in further experiments to help users to overcome their production difficulties.

- "*I think this tool could be useful to improve my pronunciation with more sounds*".
- "*I realized I cannot produce correctly similar words*".
- "*The answer given by the tool in each utterance was very fast*".
- "*I would like to produce sentences instead of isolated words*".
- "*The TTS system helped me to produce better the sounds*".
- "*I would like to hear and compare myself utterances with the system*".
- "*I would appreciate an indicator about the percentage of production success per word*".
- "*I felt frustrated after failing consecutively*".

A summary about user's opinions gathered in the session about possible CAPT GUI's improvements is:

- "*I would like to see the word's phonetic transcription beside the orthographic one*".
- "*I would like to know which words are the most difficult by assigning to them a representative color. For example, green, yellow and red, from the easiest to the most difficult ones*".

- "*I felt frustrated when the timer continued counting while the ASR system was evaluating my utterance*".
- "*I would like to pause the activity to take a break*".

Participants also reported several opinions about the training dynamics of this experiment and future work, from including more game elements and playing with other people outside the class, to combining more pedagogical and feedback resources:

- "*We think these pronunciation activities can be performed in a game for smart devices*".
- "*A leaderboard would motivate myself to keep on training*".
- "*I would prefer to train and play at home*".
- "*I would like to challenge my friends in a mobile application either in the same room (local) or online*".
- "*I felt myself frustrated to be under pressure in a class with an instructor. I would prefer to train in a stress-free place*".
- "*I would appreciate a training tutorial and pronunciation instructions with mouth pictures*".
- "*I would like more training activities. For example, to be able to listen to a word of a minimal pair and select which one is the correct answer*".
- "*The level of difficulty should be adjustable to the necessities of each one*".
- "*I would like to practice either isolated words and minimal pairs*".

Finally, the instructor of the session asked to the participants if they would prefer a tool for training, for playing or for both options. Gathered results confirmed that a 80% of learners would use the system for training, a 20% would use the system for playing, and the 100% would use the application for learning through playing.

7.3 Non-guided Learning Experiment

The second practical approach of this thesis was the experiment called *Non-guided Learning*. Different pronunciation training activities were carried out when using general-purpose speech technology in the prototype developed for this experiment, *TipTopTalk!* In particular, this prototype included an implicit competition (see its definition at Section 5.1) in which users trained individually their L2 pronunciation with a gamified CAPT system for smart devices. Subjects were university students who participated voluntarily in the experiment. They practiced anytime anywhere, choosing the training activities at free will. The goal of this experiment aimed at assessing the possible learners' pronunciation improvement along time while keeping users motivated at the same time they were training. User's interaction data was automatically monitored to obtain pronunciation assessment results.

Initially, this experiment was intended for native Spanish speakers who study American English as L2. However, due to the collaboration success achieved with different academic institutions and research groups, the prototype developed for this

experiment is still active today not only with the mentioned languages but also with simplified Chinese, Spanish, Portuguese (European and Brazilian), and German[2].

7.3.1 Experimental Procedure

A one-month protocol was established for this second experiment as shown in Figure 7.5. First, the enrollment campaign lasted for 6 days. Then, the competition was active during 24 days. Subjects could take part in the competition anytime anywhere with their own devices during the protocol's interval dates (see Section 7.3.3 for more details about participants). At the end of the competition users were invited to take an optional online questionnaire of UX about the CAPT system.

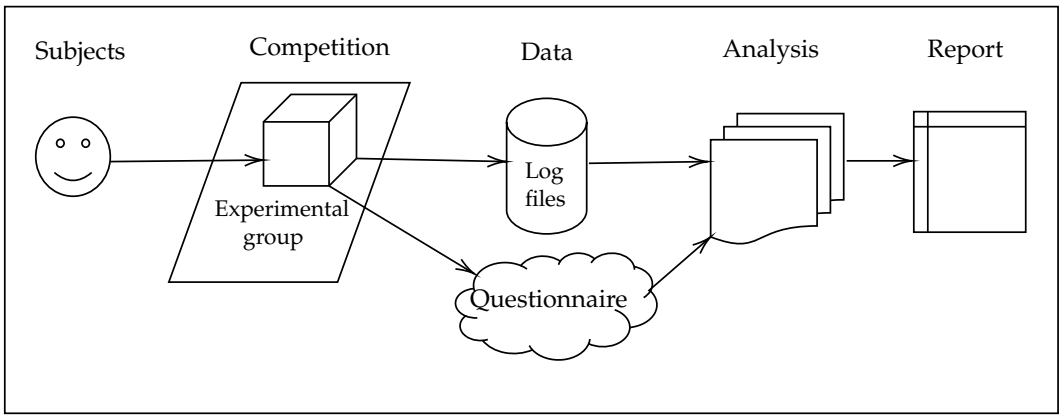

FIGURE 7.5: Steps of the TipTopTalk! prototype's protocol.

7.3.2 Enrollment

The participants' recruitment process was made by sending invitation emails to the potential participants and by means of invitation talks in selected classrooms. Students were asked to fill in a registration form with their demographic information and signed an informed consent. After registering, they received the instructions to download and install the software application from Google Play. A time window with a starting and ending date was established for a total of 24 playing days of competition. A diploma and a reward were given to those students who reached the 15 highest positions on the leaderboard. Besides, during the whole competition, the research team was available to help solving potential technical problems with the application and answered questions about its usage.

7.3.3 Participants

In this experiment, two main groups of users could be differentiated according to their L1 and L2 (see Table B.7 for specific details):

1. **Group I**: 52 native Spanish speakers between 18 and 26 years old from University of Valladolid.
   (a) **Students of English as L2**: 21 were women and 18 were men. They declare different grades of English proficiency level.

[2] https://play.google.com/store/apps/details?id=uva.eca.simm.tiptoptalk

(a) **Students of Chinese as L2**: 5 were women and 8 were men with a low Chinese proficiency level.

2. **Group II**: 6 native Chinese learners of Spanish as L2 between 18 and 26 years old from the Language Center of the University of Valladolid.

   (a) **Natives**: 5 were women and 1 was a man. They were intended to test the tool and to find possible problems or incompatibilities with the Chinese language.

   (a) **Students of English as L2**: 1 woman and 1 man with a low English proficiency level.

7.3.4 TipTopTalk! CAPT System Description

A gamified CAPT system for smart devices, called *TipTopTalk!* was developed for running the competition. It was an online game in which users played individually and their results were reflected on a general leaderboard. The main goal of a user playing with the CAPT tool was to achieve points by performing different pronunciation **activities** in **matches**, trying to reach the best position possible on a **leaderboard**. Thus, subjects learned while they were playing.

A match can be played in two modes: **Playing** and **Training**. In the Playing mode, users obtained points by performing activities based on the NCM in matches. These activities can be either discrimination, production or both of them (see their description in Section 6.2). In the Training mode, matches had also an individualistic approach. They included exposure, discrimination, or production activities. However, in this mode users did not get explicit reward nor points.

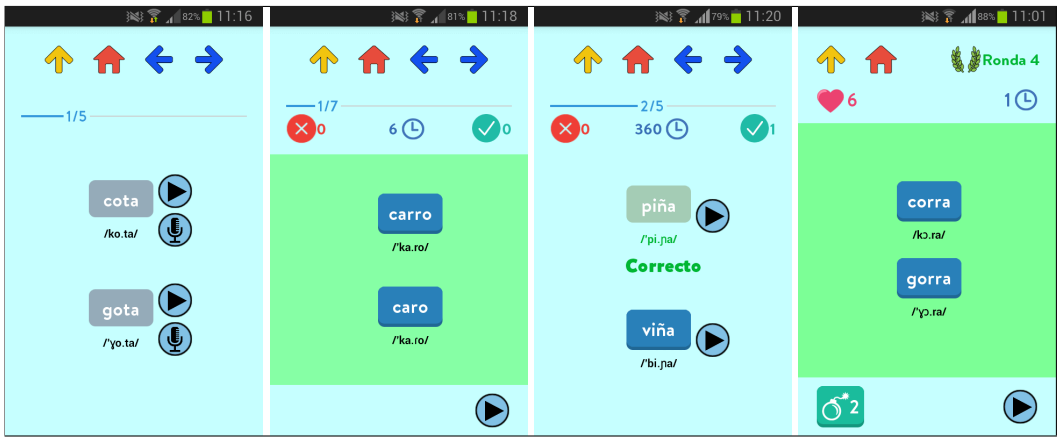

FIGURE 7.6: TipTopTalk! CAPT tool screenshots of exposure (first picture), discrimination (second picture), production (third picture), and mixed activities (fourth picture).

Participants had the freedom to select the activity types and the phonemes they want to practice. That is, users could perform a match of discrimination, of production activities, or both (see the second, third and fourth screenshots of Figure 7.6, respectively), selecting the minimal pair contrast that they want to practice. However, the system recommended the next activity mode according to the users' results. It is up to the users whether choose or not the proposed activity. In the Training mode, in addition to the Playing mode activities, discrimination and production, exposure activities were also possible to be performed (see the first screenshot of Figure 7.6).

In this experiment the free versions of the Google's off-the-self ASR and TTS systems were also included.

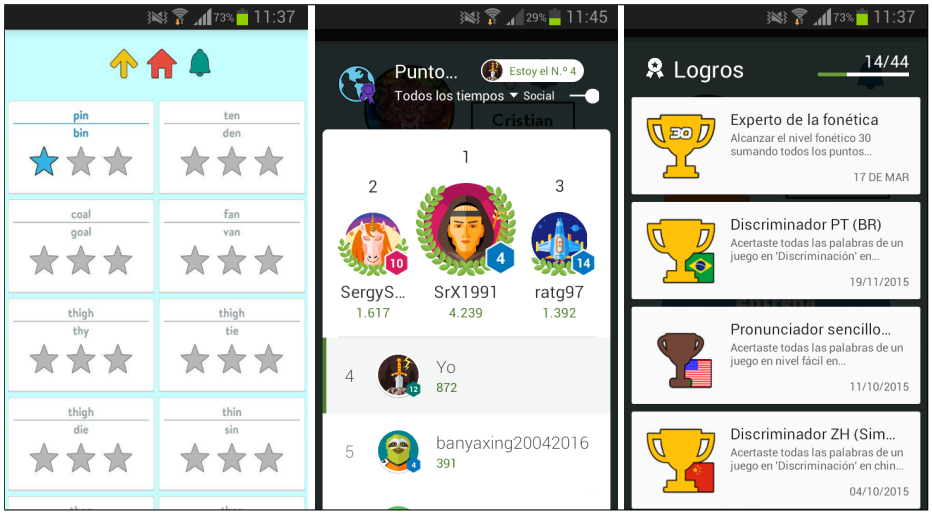

FIGURE 7.7: TipTopTalk! CAPT system screenshots of the minimal pair lists selection (first picture), the main leaderboard (second picture), and the list of achieved trophies (third picture), adapted from [19].

Furthermore, several gamification elements have been included in this CAPT system. Each activity in the Playing mode provides points to the users when they were correctly performed. These points increment the user's "phonetic level" of the game. They also help to unlock new minimal pair contrasts (see the first screenshot of Figure 7.7) and to achieve several trophies according to the activity type and difficulty level (see the second screenshot of Figure 7.7). The quantity of points depends on the activity difficulty level chosen by the user. A production attempt was considered correct (right) when the orthographic transcription of the word (or some homophone) is included in one of the first three, two, or first positions of the text hypotheses of the ASR result in the easy, medium, hard difficulty level, respectively.

Subjects improved their position on the main leaderboard of the game by accumulating points with each match (see the third screenshot of Figure 7.7). There are also different language-dependent leaderboards, based on scores attained and the number of completed rounds, where all players are ranked to increase motivation through competition.

When the difficulty level is increased, as part of the gamification strategy:

1. In discrimination activities users were asked to choose the word which has not been synthesized, rather than the synthesized one.
2. In all training mode activities the phonetic transcription of each word was hidden.
3. The remaining time per minimal pair was decreased.
4. A number of lives, which represent the number of remaining wrong attempts until finish the training mode, was included in mixed activities. In the Mixed mode users had also the possibility of skipping activities with minimal pairs when using a limited number of clear tickets.

Finally, the screen size of the user's smart device was also analyzed in order to adapt and accommodate the best experiences on a wide range of device in future experiments, since a mobile application that is optimized and responsive makes for an easier user flow and ultimately, an enjoyable experience [183].

7.3.5 Instruments

There were four different sources of data for this experiment:

- **Registration forms**: user's demographic information, such as name, age, gender, L1, academic level, and final consent to analyze all gathered data. This information was carefully collected and saved into digital text documents.
- **User's interaction log files**. The CAPT tool gathered data associated with all low-level interaction events and monitored all user activities (see Section 7.3.6 for specific details of the metrics). This data was saved into local log files and automatically uploaded to a web server.
- **Questionnaire**. An online questionnaire about UX with the CAPT system was sent via email to participants. Some of the close-ended questions were about the GUI of the CAPT system and other were specific about the training activities methodology. There was also a final open-ended question about proposals, improvements and suggestions. This data was collected and saved into a secure web server.

7.3.6 Metrics

A set of experimental variables was computed from the user's interaction log files:

1. **Training intensity**. Which computed the amount of events tracked in each session of the experiment. It consisted of number of exposure, discrimination, and recording/production activities; number of times a particular phoneme was practiced; number of attempts in each mode; and number of times a word was listened to. In particular, the number of times an activity type was performed by a user u, which is defined as:

$$A_u = \sum_{\substack{i=1 \\ Training}}^{3} A_{u,i} + \sum_{\substack{j=1 \\ Playing}}^{3} A_{u,j} \tag{7.1}$$

where A_u represents the number of times a user u performs an activity i (exposure, discrimination, and production in the Training mode) or j (discrimination, production and mixed activities in the Playing mode) in the CAPT tool. Finally, discrimination and production events, $D = \cup_u \cup_k D_{u,k}$ and $P = \cup_u \cup_k P_{u,k}$, respectively, where k includes both training i and playing j events, belong to subsets of interaction records gathered within the log files with the CAPT tool:

$$R = D \cup P \cup E \cup O \tag{7.2}$$

where R is the set of all user's interaction records gathered within the log files. D represents the amount of entries related to discrimination activities, P stands for those corresponding to production exercises E for user's exposures, and O

for other interaction events, such as activity transitions, logging in or out of the system, among others.

2. **Game-related participation**. Which considered the points and leaderboard position reached by the user; the number of trophies achieved; the number of minimal pairs lists unlocked; and the number of lives and clear tickets spent in mixed activities.

3. **Training performance**. Which measured the success attained by the participant during a specific time of each event tracked. The variable encompassed right and wrong discrimination tasks; right and wrong pronunciation tasks (with the n–best list of hypotheses and g–score values, see Section 6.2.4); success rates in discrimination and production tasks per phoneme; number of modes passed and failed; and time spent on performing events and modes.

4. **Pronunciation improvement**. Which considered the scores achieved in each training task and mode. Discrimination and production success rate values were analyzed in this experiment via two functions of quality, f_D and f_P, respectively. The contrast between the value of both quality functions f_D and f_P at a given s, relative to their initial value ($s = 0$), revealed the user's performance progression in both activities for each minimal pairs sounds. In particular, $f_D(D_{u,k}, w, s)$, computes the average number of correct answers obtained within a window of w attempts in $D_{u,k}$, beginning at the position $s = (1..N_{u,k} - w)$. $D_{u,k}$ stands for a sequence of chronologically ordered discrimination attempts by the user $u = 1..U$ of the words of a kind of pair $k = 1..K$, so that, $D_{u,k} = (d_1..d_{N_{u,k}})$, where $N_{u,k}$ represents the number of times a user u tries to discriminate words of a kind of pair k. In a similar way, production quality is defined as $f_P(P_{u,k}, w, s)$. It measures the quality of pronunciation attempts of a user u in relation to the words of a kind of pair k within a window of w words beginning at position $s = (1..M_{u,k} - w)$. $P_{u,k}$ represents a sequence of chronologically ordered production attempts by the user $u = 1..U$ of the words of a kind of pair $k = 1..K$, so that: $P_{u,k} = (p_1..p_{M_{u,k}})$, where p_i represents the attempts to produce words of a kind of pair k and $M_{u,k}$ stands for the number times that user u tries to produce words of a kind of pair k.

7.3.7 Results

The results obtained in this experiment along to those derived from the first one, Alpha, reinforced the answers to the research question RQ1 (and *Issue 1.1*) of this thesis. Besides, they tried to give answer to the research questions about the implications of the training activities and gamification elements included in the CAPT system on user's performance and motivation (RQ2 and RQ3), by following the steps defined in the research objectives RO1, RO2, RO3, and RO4.

Three different categories of results can be differentiated. First, results of users behavior when using the system (i.e., participation and type of activities selected). Second, learner's performance while interacting with the CAPT system during the competition. Finally, results from the questionnaire provided at the end of the experiment. The most important results are reported in the following subsections and discussed in Chapter 8. Performance-related results have been partially published in [18], [19], [20]; whereas results related to the gamification elements included in the CAPT system have been published in [21], [22].

User's Behavior

First, an analysis about user's behavior with the CAPT system was carried out since the selection of activities was left to user's choice. As explained in Section 7.3.4, learners could train their L2 pronunciation with six different activity types, three in the Playing mode and three in the Training one. Table 7.8 shows the activity type selected by a user (columns), depending on the previous one (rows). Discrimination activities of the Playing mode (*DIS-P*) were the most performed activity since they represented a 60.67% of the user's choice (1791 times). They were followed by *MIX-P* and *PRO-P* activities, also of the Playing mode (20.66% and 8.97%, 610 and 265, respectively). The three least performed activity types were the three Training mode ones which altogether did not reach a 10% (101, 136, and 49 times, respectively).

	Activity Type						
Previous Activity Type	***EXP-T***	***DIS-T***	***PRO-T***	***DIS-P***	***PRO-P***	***MIX-P***	
None	27	3	4	11	3	6	**54**
EXP-T	32	31	5	28	1	1	**98**
DIS-T	14	69	19	18	2	3	**125**
PRO-T	0	13	15	8	3	5	**44**
DIS-P	18	11	4	1627	53	59	**1772**
PRO-P	3	4	0	35	189	29	**260**
MIX-P	7	5	2	64	14	507	**599**
	101	**136**	**49**	**1791**	**265**	**610**	

TABLE 7.8: User's behavior according to the number of times an activity type performed in the TipTopTalk! prototype. The *None* row refers to the activities performed after the installation of the CAPT tool in the user's smart device. *EXP-T*, *DIS-T*, and *PRO-T* were exposure, discrimination, and pronunciation activity types of the Training mode, respectively. *DIS-P*, *PRO-P*, and *MIX-P* stand for discrimination, pronunciation, and mixed activity types of the Playing mode, respectively.

Table 7.8 also confirms the *DIS-P* activities as the most selected type after performing the same activity type (55% of the total, 1627 times). The second and third most performed activity types certified that users tended to repeat the same activity (*MIX-P* and *PRO-P*, 507 and 189 times, respectively). Finally, the preferred activity types performed by a user who installed the CAPT software for the first time were the *EXP-T* ones (50% of the total, 27 times).

Second, in Table 7.9 the average time spent by the participants in each activity type categorized by their gender, L1 and L2, is represented. The most performed activity type by users reported in Table 7.8, *DIS-P*, and the same activity in the Training mode, *DIS-T*, were the fastest activities carried out (34s and 66.75s for *DIS-T* and *DIS-P*, respectively, in en_EN as L2; and 31.25s and 33.5s for *DIS-T* and *DIS-P*, respectively, in cn_ZH as L2). Furthermore, the least performed activity type of the Training mode, *PRO-T* (136 times) and of the Playing mode, *PRO-P* (265 times), were the slowest performed activity types (352.5s and 306.25s for *PRO-T* and *PRO-P*,

L2	en_EN						
L1	es_ES		cn_ZH		ANY		
Gender	F	M	F	M	F	M	ANY
#Subjects	18	21	1	1	19	22	41
Training mode							
EXP-T	94	97	-	-	94	97	95.5
DIS-T	34	39	-	29	34	34	34
PRO-T	166	539	-	-	166	539	352.5
Playing mode							
DIS-P	31	100	36	-	33.5	100	66.75
PRO-P	228	456	85	-	156.5	456	306.25
MIX-P	81	132	-	-	81	132	106.5

L2	cn_ZH						
L1	es_ES		cn_ZH		ANY		
Gender	F	M	F	M	F	M	ANY
#Subjects	5	8	5	1	10	9	19
Training mode							
EXP-T	148	114	80	58	114	86	100
DIS-T	44	42	27	12	35.5	27	31.25
PRO-T	103	230	30	-	66.5	230	148.25
Playing mode							
DIS-P	40	32	26	36	33	34	33.5
PRO-P	223	252	107	87	165	169.5	167.25
MIX-P	-	-	119	108	119	108	113.5

TABLE 7.9: Average time (s) spent by users in each activity type of the TipTopTalk! prototype. The left tabular refers to American English as L2 subjects and the right one to Simplified Chinese as L2 ones. *F* and *M* mean 'female' and 'male', respectively. *EXP-T*, *DIS-T*, and *PRO-T* were exposure, discrimination, and pronunciation activity types of the Training mode, respectively. *DIS-P*, *PRO-P*, and *MIX-P* stand for discrimination, pronunciation, and mixed activity types of the Playing mode, respectively.

respectively, in en_EN as L2; and 148.25s and 149.75s for *PRO-T* and *PRO-P*, respectively, in cn_ZH as L2).

Male native Spanish participants spent more time performing *PRO-T* and *PRO-P* activities than female ones, this difference being higher in *PRO-T* in both L2 cases (539s vs. 166s, en_EN as L2, respectively; and 230s vs. 103s, cn_ZH as L2, respectively); and in en_EN *PRO-P* activities (456s vs. 228s, respectively). Finally, *PRO-T* and *PRO-P* activities were performed faster in cn_ZH as L2 than in en_EN as L2 (352.5s vs. 148.25s, *PRO-T*, respectively; and 306.25s vs. 167.25s, *PRO-P*, respectively).

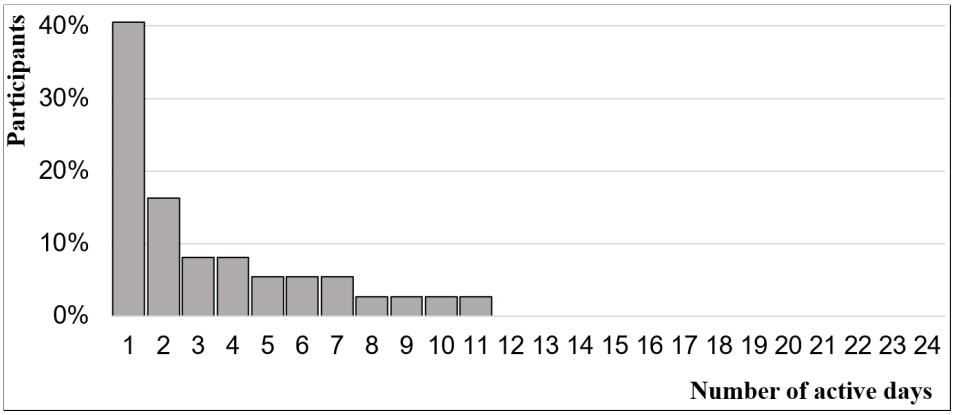

FIGURE 7.8: Distribution of users by number of days with active participation in the competition of the TipTopTalk! prototype.

Concerning the distribution of players activity throughout the 24 competition

days, Figure 7.8 shows the accumulative number of days in which a player performed at least one activity type. A high number of users only played with the game occasionally (41% participated one day). Regarding the rest of the users, the plot describes that there was not a single user who participated more than 11 days.

User's Performance

In this experiment a database of 87,918 entries was stored containing all user's interaction data with the CAPT system. In particular, almost a 40% of these entries were related to the two different event types which lead to achieve points for the competition: perception and production activities. The former events were performed by learners in discrimination and mixed activity types of the Playing mode and in discrimination activities of the Training mode. Production events were completed in production and mixed activity types of the Playing mode and in production activities of the Training one. Table 7.10 shows the average number of these two activity events performed by each user in the experiment.

Discrimination events were performed by more users than production ones in both the Training (86% vs. 62%) and in the Playing mode (100% vs. 64%). Although the selection of activity types was left to user's will, results reveal a balanced choice between perception and production activities since there were no statistically significant differences between them in the average number of events performed in each mode (405.2 vs. 349.9 in the Playing mode; and 37.0 vs. 24.3 in the Training mode). The higher average number of Playing activities than Training ones performed by each user leads to statistically significant differences (U = 442.0, $p < 0.001$, Mann–Whitney U test [184]).

	$\overline{\#Events}$	$\#Participants \vert \#Total$
Training mode		
Discrimination	37.0 (60.4%)	25/29 (86%)
Production	24.3 (39.6%)	18/29 (62%)
Playing mode		
Discrimination	405.2 (53.7%)	36/36 (100%)
Production	349.9 (46.3%)	23/36 (64%)

TABLE 7.10: Average number of discrimination and production events per participant of the TipTopTalk! prototype. The third column ($\#Participants \vert \#Total$) refers to the number of subjects who perform these activities (first value) and the total number of participants who perform an activity of the same mode (second value).

Figure 7.9 represents the number of discrimination and production activities performed by learners per competition day. The highest value of activity was reached during the middle days of the experiment (42% and 50% in discrimination and production activities, respectively). A large number of discrimination activities were carried out during the first ten days of competition (65%). In the case of production activities, a 70% of the total events were registered from the 10th day of competition. Finally, a peak of discrimination activities was observed on the penultimate day of competition (7.6%).

These perception and production events performed in the competition are represented in Figure 7.10. It shows the evolution of quality functions f_D and f_P along a

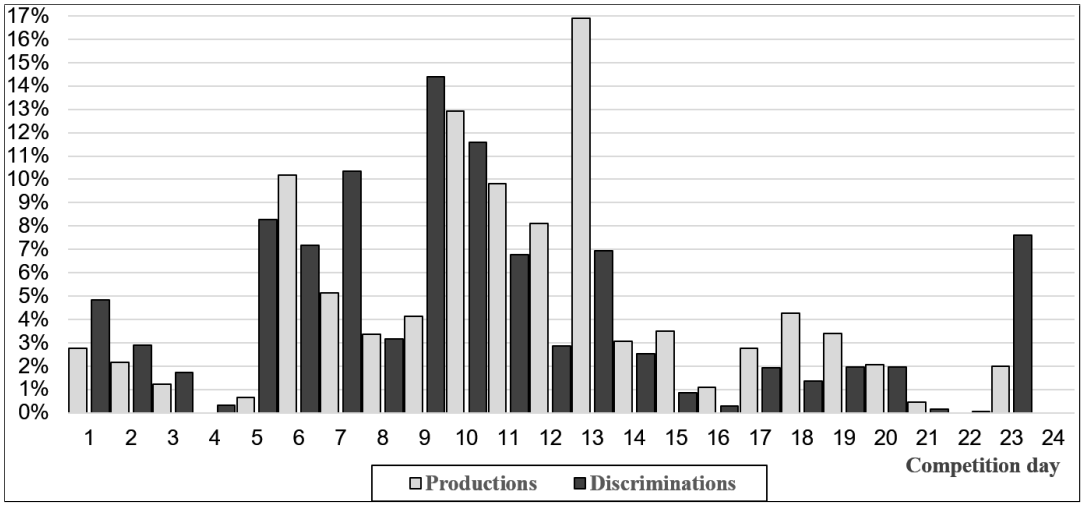

FIGURE 7.9: Distribution of discrimination and production activities per day in the TipTopTalk! prototype.

chronologically ordered attempts sequence, s, varying u and k with a window size of $w = 6$. (see Section 7.3.5 for their definition details). Three groups of users are displayed depending on the value of the quality functions achieved in the introductory window $s = 6$, which represents the initial competence of each user before using the CAPT system for the first time.

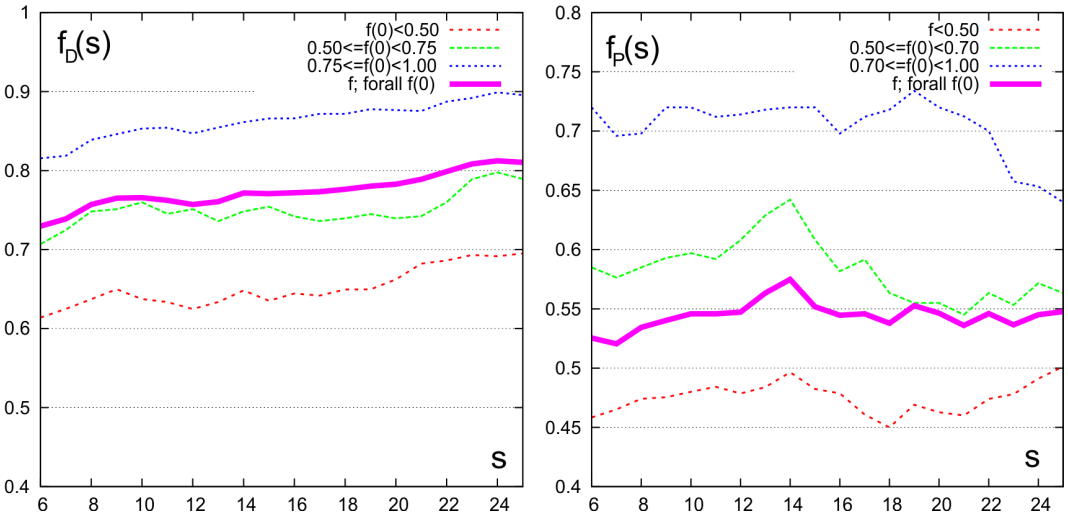

FIGURE 7.10: Evolution along time of the pronunciation quality functions of the TipTopTalk! prototype, adapted from [18]. The first diagram refers to perception activities and the second one to production activities. The ordered sequence of minimal pairs attempts was represented in the abscissa and the quality function in the ordinate axis.

Production activities registered different results. First, a positive tendency was displayed for the lowest initial value group of users (0.460 to 0.485) and the intermediate ones (0.585 to 0.645) according to their f_P, until $s = 14$. Then, these values varied, reaching a higher value (0.495) than the initial one (0.460) in the worst group and a lower value (0.570) respecting the initial one (0.585) in the intermediate group. Finally, subjects with the highest $f_P(6) = 0.730$ gravitated around this value until $s = 20$, in which f_P fell to 0.655.

Table 7.11 displays the users' success rate average in each activity type, categorized by their gender, L1 and L2. In relation to Figure 7.10, learners achieved the worst success rate values in all production activity types (32.5% and 37.7% in *PRO-T* and *PRO-P*, en_EN as L2, respectively; and 34.5% and 41.9%, in *PRO-T* and *PRO-P*, cn_ZH as L2, respectively). Discrimination results in both L2 targets confirms the improvement detected in the sequence of events along time represented in Figure 7.10, reaching higher success rate values than 60% in all cases except in *DIS-T* of en_EN as L2 (59.2%).

Besides, users obtained better success rate values in all activity types of the Playing mode than in the Training mode. Female subjects achieved better results than male ones in all cases (*F* and *M* columns of the *L1-ANY* row). These results conformed to those reported in Table 7.9, in which female participants spent less time to perform the activities than male ones. Furthermore, native Chinese speakers had difficulties with perception activities since they did not reach success values higher than 50%.

L2	**en_EN**						
L1	**es_ES**		**cn_ZH**		**ANY**		
Gender	**F**	**M**	**F**	**M**	**F**	**M**	**ANY**
#Subjects	18	21	3	1	21	22	43
Training mode							
DIS-T *(%)*	67.4	65.9	-	44.4	67.4	55.2	59.2
PRO-T *(%)*	42.7	22.2	-	-	42.7	22.2	32.5
Playing mode							
DIS-P *(%)*	76.8	76.4	82.3	-	79.6	76.4	78.5
PRO-P *(%)*	36.8	35.0	41.2	-	39.0	35.0	37.7
MIX-P *(%)*	62.6	62.5	-	-	62.6	62.5	62.6

L2	**cn_ZH**						
L1	**es_ES**		**cn_ZH**		**ANY**		
Gender	**F**	**M**	**F**	**M**	**F**	**M**	**ANY**
#Subjects	5	8	5	1	10	9	19
Training mode							
DIS-T *(%)*	60.0	78.3	70.8	50.0	65.4	64.1	64.8
PRO-T *(%)*	18.9	-	50.0	-	34.5	-	34.5
Playing mode							
DIS-P *(%)*	68.8	70.4	91.8	33.3	80.3	51.9	66.1
PRO-P *(%)*	24.3	39.3	62.1	-	43.2	39.3	41.9
MIX-P *(%)*	-	-	84.2	66.7	84.2	66.7	75.4

TABLE 7.11: Success rate (%) in each activity type of the TipTopTalk! prototype. The left tabular refers to American English as L2 subjects and the right one to Simplified Chinese as L2 ones. *F* and *M* refer to 'female' and 'male', respectively. DIS-T and PRO-T are discrimination and pronunciation activity types of the Training mode, respectively. DIS-P, PRO-P, and MIX-P stand for discrimination, pronunciation and mixed activity types of the Playing mode, respectively.

Terminal Characteristics

User's preferences about the device where the CAPT system was installed were also analyzed for future GUI improvements. Data gathered from participants who gave permission to share their smart device's technical specifications led to three main groups of Android devices in this experiment. In particular, there were 45 smart devices with a screen size lower or equal than 5.5 inches (70.4%), 8 devices with a screen between 5.5 and 7 inches (12.4%), and 11 devices (17.2%) with a size equal or higher than seven inches.

Questionnaire

In this section we present the results obtained from the voluntarily-answered questionnaire about the CAPT system UX. Regarding the Likert-type scale questions (Figure 7.11), a 90% of the questionnaire respondents did not find difficulties or consider them insignificant when interacting with the system (question 1); while the same percentage thought the guide texts and tips helped them properly to understand the proposed activities (question 2). These results were in tune with the 80% of users who disagreed or fully disagreed in feeling lost to continue (question 3). In this case, they agreed and fully agreed that the support system, its controls, and commands were adequate and useful (80% and 90%, respectively, (question 4 and question 5). A 80% of the questionnaire respondents felt confident using the system (question 6); whereas a 50% reported frustration in at least one occasion and the other 50% did not, which confirmed the lower production success rate values shown in Table 7.11 (question 7). These results were reinforced with the fact that the 95% of users agreed and fully agreed in founding the mechanics of the system easy to understand (question 9). Finally, a 30% and a 45% of the questionnaire respondents agreed and fully agreed with finding fun the game activities, respectively (question 8).

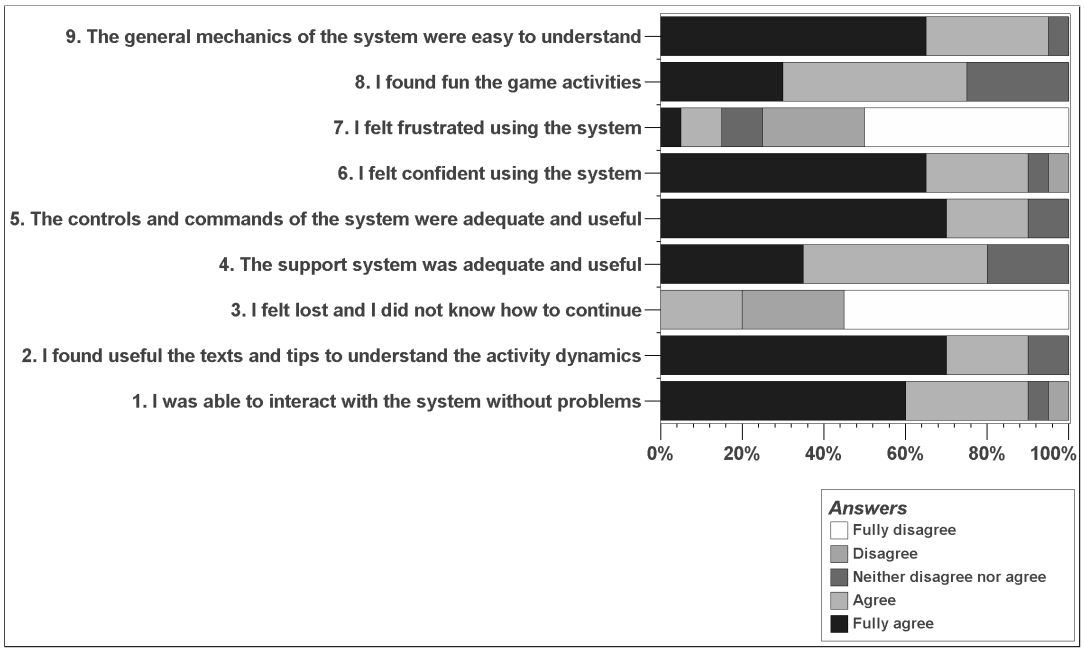

FIGURE 7.11: Likert scale questions of the TipTopTalk! prototype.

Figure 7.12 displays the answers to the last five questions of the questionnaire about selection. Over half of the answers claim *Discrimination* as the favorite training activity type (question 1). However, the *Mixed* activities were preferred in the Playing mode confirming the results presented in Table 7.8 (question 2). Besides, the 85% of the questionnaire respondents claim that the words' difficulty was adequate (question 3). Finally, up to a 90% of the users would like to challenge other people with the activities of the system and elaborate their own list of words (question 4 and question 5).

Finally, the answers provided to the optional open-ended question, "*Please, tell us any other suggestion, complaint, or future improvement*", were:

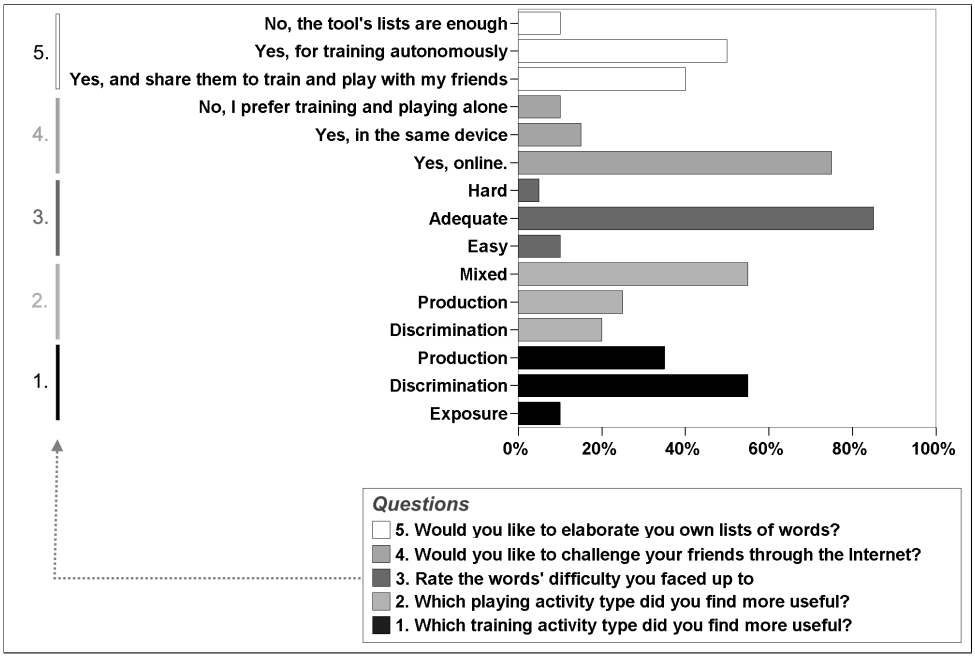

FIGURE 7.12: Selection-type questions of the TipTopTalk! prototype.

- "*I would like to have my own lists of words to practice at any time*".
- "*I would like some type of daily rewards such as points or lives*".
- "*I would appreciate the translation into my native language Chinese words*".
- "*I felt frustrated when I saw a lot of wrong production attempts*".
- "*Whenever I rotate my phone's screen, the 'success' sound is played again*".
- "*I would prefer not to scroll in the menus of the application*".
- "*Whenever I rotate my phone's screen, the activity tip disappears*".
- "*At the end of an activity my smartphone's back button is disabled*".

Some of the answers agreed with the close-ended questions (i.e., elaborating own lists of words and feeling frustration after several wrong production attempts); whereas others recommended proposal and future improvements (i.e., extra rewards and word translations). The rest of the answers were related to usability aspects, such as GUI improvements and software bugs detected.

7.4 Guided Learning Experiment

In the third experiment conducted in this dissertation, named *Guided Learning*, a controlled training protocol with a pre/post test strategy was followed to ascertain the pronunciation level improvement of the participants from different training groups. During the days between the tests, some participants trained with a pedagogical evolved version of the CAPT system from the previous experiments, which followed a specific training protocol. Two prototypes were developed for this experiment, corresponding to different L1 and L2. Native Spanish learners of English

participated in the first one, ***English Vowels***, while native Japanese learners of Spanish were involved in the second prototype, ***Japañol***. Next subsections describe their specific details.

7.4.1 Experimental Procedure

A four-week protocol was defined for this experiment. It included a pre-test, three training sessions, and a post-test, as shown in Figure 7.13. At the beginning, the subjects took part in the pre-test session individually in a quiet testing room while the sound of the session was recorded with a microphone and an audio recorder. All the students took the pre-test under the sole supervision of a member of the research team. In the case of the English Vowels prototype, learners were asked to read aloud the 25 minimal pairs contrasts administered via a sheet of paper with no time limitation. They were free to repeat each contrast as many times as they wanted if they thought they might have mispronounced them. In particular, the test included contrasts of the English pure vowels /ɑː/, /ʌ/, and /æ/, that are usually reduced to Spanish vowel {a} [185]; vowel /e/, that is usually realized as a closer Spanish {e} [185], and vowels /iː/ and /ɪ/, often reduced by Spanish speakers to {i} [74], [186]. Readers can find more details about this test in Table C.1 in Appendix C.

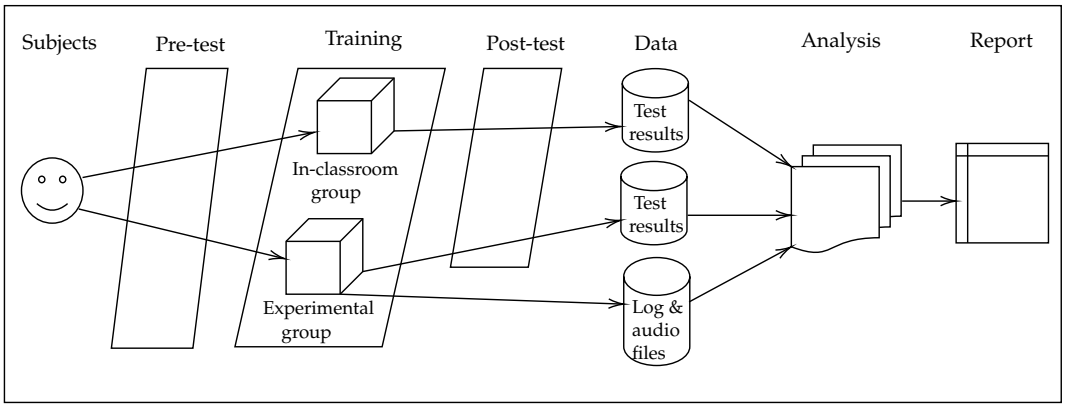

FIGURE 7.13: Steps of the English Vowels prototype's protocol, adapted from [23].

A total of three training sessions were carried out one week after running the pre-test. There was a time gap of at least 72 hours between them in order to avoid fatigue and promote memory consolidation [187]. Subjects were divided into two different groups after taking the pre-test (see Section 7.4.3 for more details). The training sessions of both groups were conducted at the same time in different locations (classroom and laboratory). A maximum time of 60 minutes was established in each one of them. The session's learning content was divided into lessons. Two lessons were presented in each session and a minimal pair contrast was practiced in each lesson (block distribution). However, most phonemes were retaken in later sessions (spaced distribution). In particular, in the first session, phonemes /ɑː/–/æ/ and /æ/–/ʌ/ were contrasted. In the second one, /ɑː/–/ʌ/ and /e/–/æ/. The last session involved the phonemes /ɪ/–/iː/ and /ɪ/–/e/. Only /iː/, a vowel that is almost interchangeable with the Spanish /i/, was left out of a repeated practice scheme.

On the one hand, students of the experimental group exclusively used the CAPT system (see Section 7.4.6 for more details). The software application was installed on an Android emulator (NOX App player) in the laboratory computers. Subjects

were inside a cubicle, separated to each other by glass dividers, and used a headset with microphone. Before starting the first session, users in the CAPT-condition were instructed in place on how to use the software. During the rest of the session, each student worked individually (see Section 7.4.5 for activity details) and did not have any interaction with either classmates or instructors. Along the three experimental sessions, a total of 72 minimal pairs were presented to participants (12 in each lesson, 2 lessons per session)[3]. On the other hand, the in-classroom group participants had interaction with their classmates and the instructor (see Section 7.4.4 for more details).

Finally, one week after the last training session, a post-test was carried out by all participants. The post-test contents and conditions were identical to those of the pre-test. Both tests were assessed by three L2–English raters experts in phonetics, in random order and independently, some days after collecting all results. The raters were specialized EFL teachers who had worked extensively on L2–English pronunciation. They had no contact with the participants. They did not know if the utterances they were evaluating belonged either to pre-test or post-test realizations.

The same number of training sessions, duration and spacing between them and the places for the English Vowels prototype (see Section 7.4.1) were repeated for the Japañol prototype. That is, a four-week protocol which included a pre-test, three training sessions, and a post-test was followed, as shown in Figure 7.14. In particular, the tests included 28 contrasts of the most difficult to produce Spanish consonant sounds by native Japanese speakers [188], [189], [190]: [θ, s] sounds are usually reduced to Japanese consonant [s]; the Spanish sounds [θ, f] are confused by Japanese speakers in perception activities since they do not exist in Japanese and the nearest Japanese sound is [ɸ]; the consonant sound [f] is often confused with [x], especially when these sounds are followed by [u]; the Spanish phoneme /r/ is usually realized as [ɾ] or [l]; and finally, the combination of fricative /f/ with /l/ or /r/ in onset, also triggers mispronunciation. Readers can find more details about this test in Table C.2 in Appendix C.

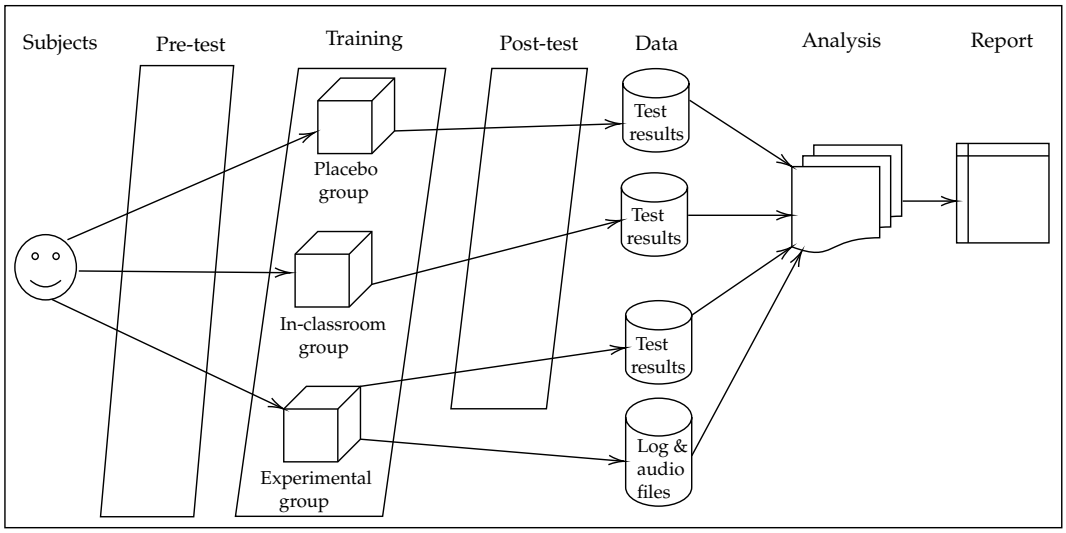

FIGURE 7.14: Steps of the Japañol prototype's protocol.

[3] https://github.com/eca-simm/minimal-pairs-enus-eses

A total of 84 minimal pairs were presented to participants (12 in each lesson, 2 lessons per session, except for the last session that included 3 lessons)[4]. A blocked and spaced practice schedule was also followed within the sessions. 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]–[fɾ], [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.

7.4.2 Enrollment

The recruitment campaign for the English Vowels prototype consisted in a call for volunteers from the same course of EFL for B1–B2 level of the Language Center of the University of Valladolid. The target subjects of the Japañol prototype were native Japanese speakers learners of Spanish from two different locations, the Language Center of the University of Valladolid and the University of Seisen (Japan). Students who gave consent, filled in a registration form with some personal information and signed an authorization. The training protocol sessions were carried out during their course's classes. All participants were awarded with a diploma and a reward after completing all stages of the experiment.

7.4.3 Participants

A total of 20 native Spanish students who qualified and registered for the same EFL course of the Language Center of the University of Valladolid were initially selected for taking part in the English Vowels prototype. This institution distributes its students along its different courses by means of an accurate level test. Two participants left the training sessions for personal reasons during the early stages, and were consequently discarded.

Before being allowed to take this course at the University, the participants took a placement test. In this case, all of them had an intermediate B1–B2 level of English with a very little or no previous training in English phonetics (see Table B.7 for specific details). In this way we ensured that the experiment realistically reproduced the diversity of students that attend the same course of the Language Center of the University of Valladolid; and that all students had the same initial level of English. Furthermore, it was explicitly requested to participants not to do any extra work in English (extra lessons, conversation exchanges with natives, etc.) while the experiment was still active.

Students were offered, through the mediation of the instructor, and with the reluctant agreement of the institution's authorities, to cover a small part of the EFL course program (in particular, the teaching of a few English phonemes) by using a CAPT system. Participants were divided into two homogeneous groups since all of them obtained low pre-test scores (see Section 7.4.12 for more details about results):

1. **Experimental group**. 10 students who trained their English pronunciation with the CAPT system developed, during three sessions of 60 minutes. 2 were women and 8 were men.

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

2. **In-classroom group**. 10 students who attended to three pronunciation teaching sessions of 60 minutes within the EFL course, with their usual instructor, making no use of any computer-assisted interactive tools. However, two of them were discarded since they left the experiment before finishing all the stages. 5 were women and 3 were men.

A total of 33 native Japanese speakers from 18 to 26 years old participated voluntarily for the Japañol prototype. All of them declared a low or intermediate level of Spanish as L2 with a very little or no previous training in Spanish phonetics (see Table B.7 for specific details). Besides, they were requested not to do any extra work in Spanish (extra phonetics research, conversation exchanges with natives, etc.) while the experiment was still active. They came from two different locations:

1. **Language Center of the University of Valladolid**. 8 students of the Spanish philology degree of the same University course who recently arrived to Spain from Japan in order to start an L2 Spanish course. 5 were women and 3 were men.

2. **University of Seisen**. 25 female students of the Spanish philology degree from Seisen, Japan.

Participants were divided into three homogeneous groups since all of them obtained low pre-test scores (see Section 7.4.13 for more details about results):

1. **Experimental group**. 18 students who trained their Spanish pronunciation with our CAPT system, during three sessions of 60 minutes. 15 are women and 3 are men.

2. **In-classroom group**. 8 female students who attended to three pronunciation teaching sessions of 60 minutes within the L2–Spanish course, with their usual instructor, making no use of any computer-assisted interactive tools.

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 *Teatro Pie Izquierdo* of Valladolid[5] (5 women and 5 men) participated in the recording of a total of 41,000 words included in the pre/post-tests and the CAPT tool developed (see details in Appendix D). The recording sessions were carried out in an anechoic chamber of the University of Valladolid. The dataset was intended to be part of an own ASR system for assessing the pre/post-test utterances gathered in the experimentation.

7.4.4 In-classroom Group Training Activities

In both prototypes, in-classroom group participants were guided by a non-native L2–English or L2–Spanish teacher, respectively, with a vast experience in phonetics of the target L2. The teaching program included the same phonemes covered in the experimental group. Each 60–minutes session began with around 10 minutes of explicit articulatory instructions and auditory descriptions of the sounds, with ample exposure to contrasting examples. These examples were both produced by the instructor and extracted from the audio materials of an English or Spanish as L2 handbook, respectively. After exposure, students were asked to practice perception

[5] http://www.pieizquierdo.es/

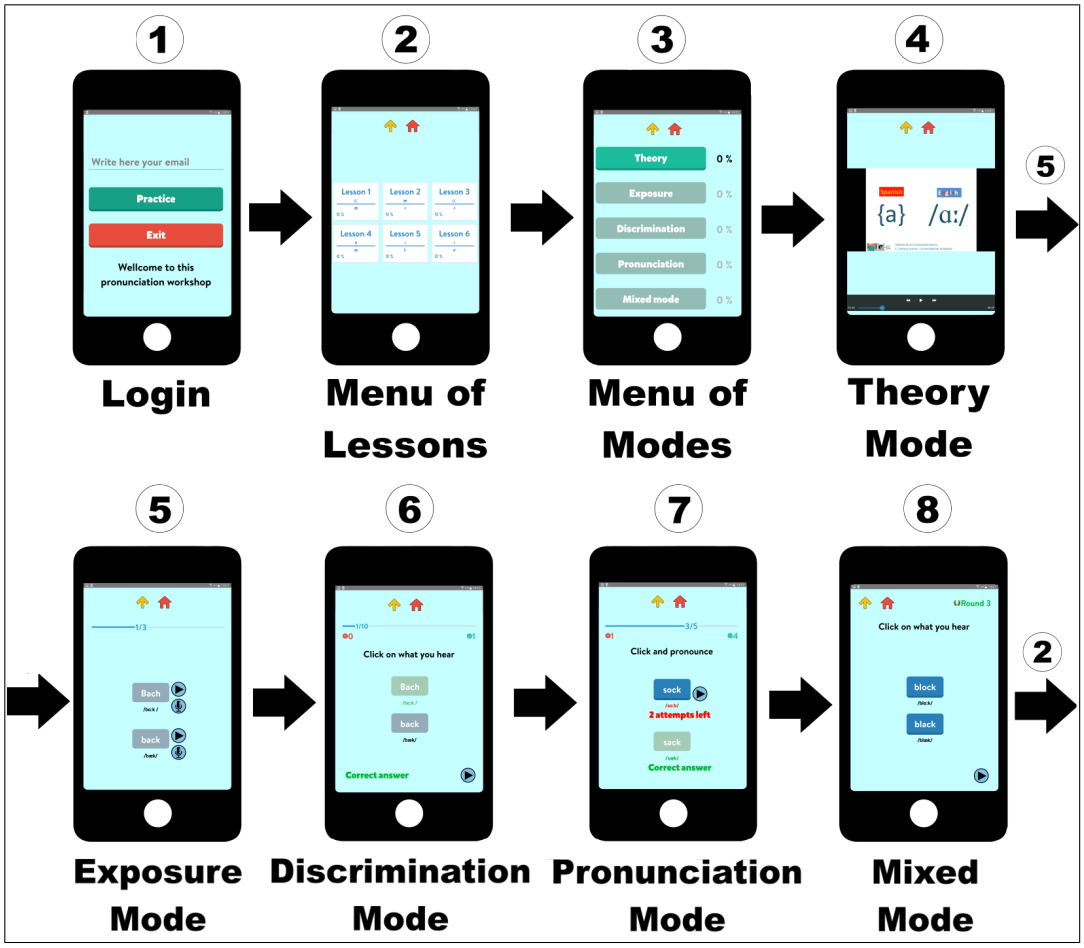

FIGURE 7.15: Standard flow to complete a lesson in the English Vowels prototype, adapted from [23].

activities using auditory materials from the same handbook. Then, learners publicly produced the sounds themselves by means of word utterances, under the supervision of the instructor. The equal participation of all the students was ensured by allotting participation turns in a uniform way. The role of the instructors was clear: they answered questions and produced model pronunciations upon request. They also diagnosed pronunciation errors publicly (that is, for the benefit of all attending students), providing the necessary corrective feedback. Each session was closed with a 5–10 minute review.

7.4.5 CAPT Tools Description

A new version of the CAPT system for both prototypes was designed and implemented, ***English Vowels*** and ***Japañol***. It consisted in a smartphone application with speech technology (Google ASR/GCSTT and Google TTS) which led users through 8 main steps or stages for each training lesson (see Figure 7.15 and Figure 7.16). The content changed according the target L1 and L2 of each prototype. The minimal pairs lists were elaborated by following the protocol defined in Section 6.1.2.

After logging in (stage 1), the user selects the lesson to be practiced (stage 2). Each lesson contains a different minimal pair contrast. The score reached by the user (expressed as a [0–100] percentage) is regularly updated on this screen. Lessons

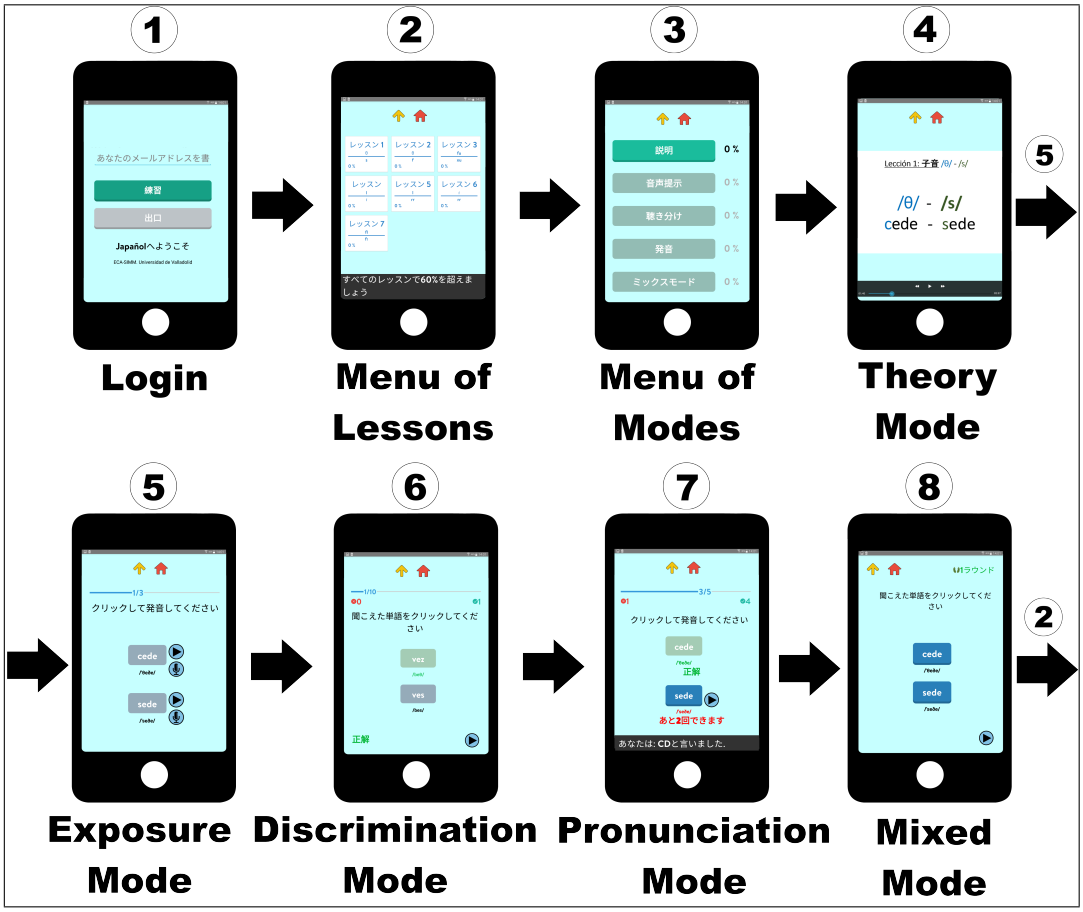

FIGURE 7.16: Standard flow to complete a lesson in the Japañol prototype, adapted from [25], [26].

must be undertaken in a consecutive order, that is, the link to lesson 2 becomes active only when lesson 1 is successfully completed, and so forth.

The sequence of training modes (stage 3) is displayed after choosing the lesson (see Section 2.1.1 for specific details about the training activities). In order to be completed, users must follow the activity flow explained in Section 7.4.6. So, each lesson takes the user through theoretical, exposure, discrimination, and production activities, *a priori*, in a strictly consecutive order. A final Mixed mode is included, to be done at the end of each lesson, where discrimination and production tasks alternate randomly. Each training mode contains a fixed number of mandatory tasks (see Table 7.12). A strict control by the system is ensured: neither lessons, nor modes within lessons, nor tasks within training modes can be skipped in any way or undertaken in any order other than the one established by the CAPT tool. Besides, learners perform the activities with the same difficulty level. Each training mode is accessible by clicking on its button on the Menu of modes, when it is enabled.

	Theory	Exposure	Discrimination	Pronunciation	Mixed
# Tasks	1	3	10	10	9

TABLE 7.12: Number of tasks of each training mode of the Guided Learning experiment, adapted from [23].

In the first training mode (stage 4), a short multimedia **video** with concepts and

tips about the articulation of sounds of the minimal pair of the lesson is displayed in the NCM fashion as explained in Section 6.2.1. The option to advance to the next mode only becomes available at the end of the video. Within the 60 minutes afforded to each session and at their own discretion, users may choose to review this material as many times as they want. At stage 5 (**Exposure** mode), the preliminary exposure to the contrasts presented in the theoretical video is reinforced. In this mode, users must listen–repeat–compare three minimal pairs wit no limit of attempts. This mode serves as a feedback recommendation by the system when users get bad results in the next training modes (see Section 7.4.6 for more details).

In the next training mode, **Discrimination** (stage 6), participants must identify the word generated by the TTS in each of the ten tasks. The number of attempts is limited to one. Users are allowed to listen to the synthesized model of the words as many times as they want, in a maximum time of 10 seconds per task. Its speed is alternated between normal and slow production rates. Stage 7 refers to the **Pronunciation** training mode. In this mode, a minimal pair for each one of the five tasks is shown. A maximum of five attempts are allowed for each word in the pair, with a time limit of 60 seconds per minimal pair. A production attempt was considered correct (right) when the orthographic transcription of the word (or some homophone) is included in the first position of the text hypotheses of the ASR result. Besides, after three consecutive failures, the system executes an explicit corrective feedback response that recommends users to listen to the synthesized version of the problematic word. In the Japañol prototype, short feedback tips after wrong production events are displayed (see Section 6.2.1). The final training mode of each lesson is the **Mixed** mode, stage 8, which works as a review mode, since it incorporates again both discrimination and production tasks. In this mode, four perception tasks and five production ones alternate randomly, summing up a total of nine task activities.

7.4.6 Experimental Group Training Activities

Both prototypes followed the same training methodology based on the NCM (see Section 6.2). In particular, the teaching program was reduced to a limited number of units (lessons), corresponding to each sound contrast (see Section 7.4.5 for more details). There were five training modes in each one (Theoretical, Exposure, Discrimination, Pronunciation, and Mixed modes). Besides, the user's freedom to select exercises at will in each lesson was sacrificed in favor of fixed and pedagogically informed training routines based on user's results. That is, the system decided on the next training mode analyzing the user's performance.

The grade of success (score) reached in each of the five training modes accumulated, in percentage terms (see Section 7.4.10), on a lesson score. The final game score (G) consisted of the average value of all lesson scores. The next lesson was available if the user attained a score of at least 60%, since a threshold over 50% reduced the incidence of success by chance, particularly in binary–choice tasks, while keeping the threshold at 60% still offers the possibility of maximally discriminating up to five levels of success (6, 7, 8, 9, 10). When this threshold was not achieved, the CAPT system suggested the user to go back to the Theory or Exposure modes before attempting the mode again, in order to review the theory of problematic vowels (Theory mode), and to perceive again (Exposure mode) the contrasting sounds practiced in the failed mode. When the review was over, users were brought back to the pending mode. Again, users could not advance to the next training mode of a lesson if they did not reach a minimum score of 60% (it was not 50% to avoid randomness).

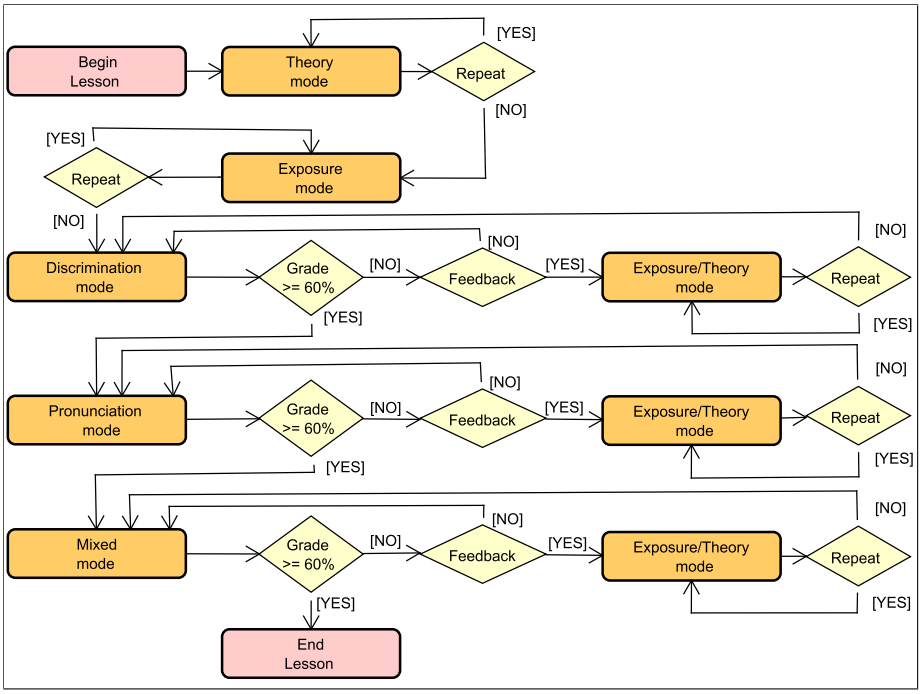

FIGURE 7.17: Guided Learning experiment training activities flowchart in each lesson, adapted from [23].

When reaching a score below 60% in any mode, users had two possibilities: repeating the training mode again, or following the training mode recommended by the system as feedback as shown in Figure 7.17.

7.4.7 Instruments

Six different sources of data were presented in both prototypes of this experiment:

- **Registration forms**: user's demographic information, such as name, age, gender, L1, academic level, and final consent to analyze all gathered data. This information was carefully collected and saved into digital text documents.
- **Pre-test**. Audio data from user's utterances at the beginning of the experiment. This data was gathered into a secure web server.
- **User's interaction log files**. The CAPT tool gathers data associated with all low-level interaction events and monitors all user activities (see Section 7.4.8 for specific details of the metrics). This data was saved into local log files and automatically uploaded to a web server.
- **CAPT system's interaction audio recordings**. Audio files from the Japañol prototype could be kept and stored since the testing days coincided with the launching of the Google's online speech API called GCSTT (see Section 6.4.1), which allows to keep the audio files after being sent to their online ASR system. In the case of the English Vowels prototype, it was not possible to store this audio data (only the n–best list and g–scores information from the recognition events). All this data was gathered into a secure web server.
- **Post-test**. Audio data from user's utterances at the end of the experiment. This data was collected and saved into a secure web server.

- **Other sources**. An online questionnaire about UX [191] and personal opinion about the CAPT system was filled in by the participants of the experimental group in both prototypes.

7.4.8 Metrics

A set of experimental variables was computed from the user's interaction log files:

1. **Training intensity**. This computed the number of events tracked in each session of the experiment. It contained the number of exposure, discrimination, and recording/production tasks; number of times a particular phoneme was practiced; number of attempts in each training mode; number of lessons and sessions in which a user has participated; and times a word was listened to (including both, listening events imposed by the system, *mandatory listening*, and those requested by the user, *requested listening*).
2. **Training performance**. For each tracked event, this metric quantified the number of right and wrong attempts during a specific time lapse. The variable encompassed right and wrong discrimination tasks; right and wrong production tasks; success rates in discrimination and production tasks per phoneme; number of training modes and lessons passed and failed; and the time spent on watching videos and performing training events, modes, and lessons.
3. **Pronunciation improvement**. This considered the scores achieved in each training task, mode and lesson. The CAPT system also provided a final software score, that is, a total score granted by the application to each user at the end of the last session (see Section 7.4.10). The experimental design included a groupwise comparison of these scores with those assigned in the human-rated post-test.

7.4.9 Subjective Perceptual Assessment

In the case of the **English Vowels prototype**, a whole **listen–and–rate procedure** was carried out for the utterances of the **pre- and post-test** by three L2–English raters experts in phonetics from the University of Valladolid. These experts could perceived slight deviations from American pronunciation and subtle processes of L1–L2 feature transfer since they have been working together for several years in the same department. They revised a mixed-up array of strictly anonymous pre-test and post-test audio files (they received no indication as to which files were pre-test and which were post-test) and applied the same criteria for scoring the tests. During the process, raters neither interacted amongst themselves nor with the subjects. They assigned scores between 1 (minimum) and 3 (maximum) to each English word produced by the participants since they can be easily entered with three fingers on a standard keyboard. When a rater was undertaken with a perfect native realization, she/he quickly entered a 3. When the pronunciation was a clearly transferred Spanish sound, the fingers did not need to travel to the zero key, but struck the nearby key 1. Intermediate scores could be entered by using decimals. Since the Spanish educational system traditionally works with a [0, 10] scale, raters frequently applied a linear scaling of the points in order to map the scores onto the traditional [0, 10] scale:

$$s_t = 5(s_i - 1); \quad s_t \in [0, 10]. \tag{7.3}$$

where s_i is the score obtained in the initial [1, 3] scale, and s_t is the score with which s_i corresponds in the target [0, 10] scale. Although segmental rating constitutes a very demanding task, this assessment procedure allowed raters for maximal concentration on sound, and minimal obtrusion of the manual mechanics of scoring. The value of each score might be described as follows:

- **Score 1**. The sound produced was indistinguishable either from an L1 sound or from another L2 sound.
- **Score 1.5**. The sound produced was still very close to the L1 sound, although it differed from it in just the right way. Or the sound produced seemed to be midway between two L2 sounds.
- **Score 2**. The sound produced was sufficiently different from any other L1 or L2 sound, and it was near the area of the L2 target sound.
- **Score 2.5**. The sound produced was clearly not an L1 sound, and it was recognizable as the target L2, although some anomalous features remained.
- **Score 3**. The L2 sound was perfectly native, and therefore totally different either from any L1 sound, or any other L2 sound.

A second rating procedure performed in the English Vowels prototype consisted on carrying out an **ABX test** [192] comparing some pre/post-test utterances. Each rating unit consisted of two versions of the same mixed minimal pair, produced by the same speaker, one from the pre-test and the other from the post-test, presented in random order. The mixed minimal pairs selected were the numbers 1, 3, 5, 6, and 11 from the pre-test and post-test (Table C.1 in Appendix C). None of the English words contained in these pairs are practiced as part of the CAPT system's program. A total of 90 rating units were assessed by the raters via a web page (5 mixed minimal pairs and 18 subjects).

	Q1	**Q2**
	In which of the two pairs (A and B) is the English word better pronounced?	***Would you say that the speaker has reached a native pronunciation on her/his best realization of the English word?***
A1	*I do not know*	*I do not know*
A2	*Pair A is better*	*Yes, it is native-like pronunciation*
A3	*Pair B is better*	*It is almost native*
A4	*There is no difference*	*Not at all*

TABLE 7.13: ABX questions and answers of the English Vowels prototype. Q1 and Q2 are the questions and A1 to A4 are the answers.

In addition to these rating units, an extra validation rating unit for each pair (5 more rating units) performed by a bilingual Spanish–English speaker was introduced in the ABX for further ascertaining the reliability of rating. He was asked to perform the pair once with a transferred pronunciation of the English item, and a second time distinguishing as clearly as possible between the Spanish and the English realizations. Both, the right and wrong pairs produced by this speaker were also presented to the raters in random order, together with the participants utterances. In the ABX procedure, six raters were asked to confront and assess each rating unit anonymously. They were also specialized pronunciation professors from

the University of Valladolid. They were informed that the first element of the pair is always the Spanish word, and the second one the English word. In each rating unit, one of the pairs was randomly tagged as A, and the other as B. The rater's task consists in confronting each rating unit and answering two questions by selecting one of the four possible answers (see Table 7.13).

Regarding the second prototype of this experiment, **Japañol**, since the number of audio samples and subjects was so much higher than the previous one, five expert phoneticians and native speakers assigned a correct/incorrect value to each **pre- and post-test** word of the participants of the Language Center of the **University of Valladolid** (see 7.4.3); whereas **the rest of utterances** of the participants from the **University of Seisen** were objectively assessed by two different **ASR systems** (see section 7.4.10). In the first case, all data was presented to human raters randomly and without user association via a web page. They were asked to focus on the specific sound of each word which should be generated correctly, ignoring the bad pronunciations of the rest of sounds. During the process, raters neither interacted amongst themselves nor with the subjects. **Experts scores** were computed by summing up the number of correct words per speaker and normalizing the result to the range [0, 10].

7.4.10 Scoring Procedures

An objective value of user's performance was defined for both prototypes of this experiment, called **game score**, G, which consisted of the average value of the success scores obtained by a user in each of the training lessons using the CAPT system:

$$G = \frac{1}{N}\sum_{i=1}^{N} L_{s,i}; \quad G \in [0, 10]. \tag{7.4}$$

where s is the speaker, i is the lesson, and N refers to the number of lessons attempted ($N = 6$ in the English Vowels prototype and $N = 7$ in the Japañol prototype). In particular, each lesson can be rated by a score, $L_{s,i}$, based on the user's performance in the Discrimination (D), Pronunciation (P), and Mixed (M) modes (see Section 7.4.5 for specific details of each one). It can be expressed as:

$$L_{s,i} = \frac{1}{3}(D_{s,i} + P_{s,i} + M_{s,i}); \quad L_{s,i} \in [0, 10]. \tag{7.5}$$

The score in the Discrimination mode, $D_{s,i}$, is based on the number of discrimination tasks (DT) successfully attempted (see Table 7.12):

$$D_{s,i} = \sum_{j=1}^{10} \mathrm{DT}_j; \quad D_{s,i} \in [0, 10]. \tag{7.6}$$

where DT_j is the discrimination task's value (1 if right, 0 if wrong), s is the speaker and i is the lesson. The score in the Pronunciation mode, $P_{s,i}$, is based on the number of production tasks (PT) successfully carried out (see Table 7.12):

$$P_{s,i} = \sum_{j=1}^{10} \mathrm{PT}_j; \quad P_{s,i} \in [0, 10]. \tag{7.7}$$

where PT_j is a production task (1 if right, 0 if wrong), s is the speaker and i is the lesson. The score in the Mixed mode is based on the number of Mixed-mode tasks (MT) successfully attempted (see Table 7.12):

$$M_{s,i} = \frac{10}{9} \sum_{j=1}^{9} \mathrm{MT}_j; \quad M_{s,i} \in [0, 10]. \tag{7.8}$$

where MT_j is a mixed task (1 if right, 0 if wrong), s is the speaker and i is the lesson.

In the case of the Japañol prototype, the second objective method for assessing user's utterances consisted on processing all utterances with GCSTT and Kaldi ASR engines, obtaining an n–best list of hypotheses and their confidence values. The **ASR score** was computed by summing up the number of correct words per speaker and normalizing the result to the range [0, 10].

7.4.11 Statistical Tests

Since most data gathered did not pass the Kolmogorov–Smirnoff nor Levene's standard tests for assuming normality and homogeneity of variances, respectively, several non-parametric tests for non-normally distributed data were carried out to detect statistically significant differences:

1. Expecting that there might be a certain degree of variability in the scoring process by human agents, a consistency check based on Kendall's coefficient [193] analysis was carried out.
2. Inter- and intra-group pairs comparisons between the pre/post-scores were carried out using a Mann–Whitney U test [184] and Wilcoxon signed-rank test [194], respectively.
3. A Mann–Whitney U test was also performed to analyze significant differences between following or not the recommended feedback.
4. A Kappa Fleiss index [195] computed the inter-rater consistency in the ABX and perceptual tests.
5. A Chi–square test [182] was run to measure the statistically significant differences between the two-way contingency table of the ABX results of the experimental and in-classroom groups.
6. A Pearson correlation [196] was carried out to compare the scores assigned by the software and the human raters scores of the post-test.

7.4.12 English Vowels Prototype Results

This prototype was the first one to focus on a guided and pedagogical protocol (see Section 7.1). It mainly tried to give answers to the research question RQ2 (and its *Issues*) about the effects on user's pronunciation improvement following a specific pedagogical training methodology, and also to reinforce the answers to RQ1 (and *Issue 1.1*) about the inclusion of current ASR and TTS systems into CAPT systems in a non-obstructive way for any L2 pronunciation proficiency level. This prototype also followed the steps defined in the research objectives RO1, RO2, RO3, and RO4.

The most relevant results obtained in this prototype are presented in next subsections according to: results gathered from the interaction between the learners and

the CAPT system during the training sessions related to performance (1) and behavior (2), and results extracted from the pre-test and post-test (3). These results have been partially published in [23], [24], and a discussion of them is addressed in Chapter 8.

User's Performance

Table 7.14 reports the intensity of use of the CAPT tool by the speakers. $\overline{n}$, m, and M are the mean, minimum and maximum values, respectively. *Time (min)* row stands for the time spent on minutes per learner in each mode in the three sessions of the experiment. *#Tries* represents the number of times a mode was executed by each user. The symbol - stands for 'not applicable'. *Mand.* and *Req.* mean mandatory and requested (listening). The TTS system was employed in both listening types; whereas the ASR was only used in the *#Productions* row.

	Theory			**Exposure**			**Discrimination**			**Pronunciation**			**Mixed**		
	$\overline{n}$	m	M	$\overline{n}$	m	M	$\overline{n}$	m	M	$\overline{n}$	m	M	$\overline{n}$	m	M
Time (min)	31.32	20.1	39.2	16.93	11.1	29.6	5.48	3.7	7	41.47	19.2	65.1	19.03	3.7	34.1
#Tries	6.4	6	8	11.9	7	17	7.2	6	9	12.6	6	21	9	6	18
#Mand.List.	-	-	-	347	210	510	69.5	60	82	-	-	-	26.8	15	54
#Req.List.	-	-	-	146.9	64	292	29.9	0	75	147.9	25	426	63.2	20	178
#Discriminations	-	-	-	-	-	-	69.5	60	82	-	-	-	26.8	15	54
#Productions	-	-	-	-	-	-	-	-	-	441.5	166	806	174.1	87	382
#Recordings	-	-	-	90.2	56	134	-	-	-	-	-	-	-	-	-

TABLE 7.14: User's performance with the CAPT system of the English Vowels prototype, adapted from [23].

A high rate of active user-time invested in interactive tasks is shown in Table 7.14 (114.0 minutes out of the total 180.0 minutes in three sessions). The activity registered with speech technology systems reached high values since as an average term, each subject listens to synthesized words for 831.2 times (calculated as the sum of the values of *#Mand.List.* of Exposure, Discrimination, Pronunciation, and Mixed modes, and *#Req.List.* values of column $\overline{n}$) and uses the ASR system 615.6 times (calculated as the sum of the value $\overline{n}$ of the column *#Productions* of Pronunciation, and Mixed modes), reaching a rate of 8.04 uses of the TTS/ASR per minute.

On the other hand, the variation in the minimal and maximum values of the variables shown in Table 7.14 clearly illustrates the differences between users. For instance, the fastest user in performing the Mixed mode's activities spent 3.7 minutes; whereas the slowest one took 34.1 minutes. This contrast can also be observed in the time spent on the rest of the training modes and in the number of times learners practiced each one of them (row *#Tries*). These inter-user differences also affected both the number of times the users made use of the TTS (109 min. vs. 971 max.) and the number of times they used the ASR (253 min. vs. 1188 max.).

The two main interactive training tasks —discrimination and production— included in Discrimination, Pronunciation, and Mixed modes, motivated the variety in the use of the tool regarding the differences between users, since a high quantity of attempts which involved them was registered. This affirmation is further illustrated in Table 7.15 where the number of correct and incorrect interactions per tested phoneme is shown. The final column (*Total*) indicates that production activities resulted tougher than discrimination ones: 53.5% vs. 81.2% of successful events,

Successful (S) and Failing (F) Events														
Task	ɑ:		æ		ʌ		e		ɪ		i:		Total	
	S (%)	F	S (%)	F	S (%)	F	S (%)	F	S (%)	F	S (%)	F	S (%)	F
Discrimination	143 (75.7%)	46	198 (81.1%)	46	114 (77.0%)	34	144 (86.2%)	23	105 (78.9%)	28	78 (95.1%)	4	782 (81.2%)	181
Production	151 (36.2%)	266	261 (53.6%)	226	127 (42.1%)	175	195 (76.5%)	60	115 (58.4%)	82	103 (85.1%)	18	952 (53.5%)	827
All productions	151 (8.9%)	1543	261 (15.5%)	1424	127 (10.6%)	1066	195 (31.1%)	433	115 (17.0%)	563	103 (37.1%)	175	952 (15.5%)	5204

Mandatory (M) and User-Requested (R) Listening Events														
	ɑ:		æ		ʌ		e		ɪ		i:		Total	
	M	R	M	R	M	R	M	R	M	R	M	R	M	R
Discrimination	189	86	244	89	148	62	167	75	133	74	82	24	963	410
Production	-	562	-	552	-	374	-	218	-	241	-	53	-	2000

TABLE 7.15: Right, wrong, and listening events categorized by phoneme of the English Vowels prototype, adapted from [23]. The symbol - stands for 'not applicable'.

respectively. This difference was accentuated when the *All productions* events rate was compared: 15.5% vs. 81.2%. As a reader's reminder, a maximum discrimination and production wrong sequence consisted in one and up to five wrong attempts, respectively.

The most difficult phonemes for the learners can also be revealed from Table 7.15, since there were important differences that affect both discrimination and production activities. For instance, the phoneme /ɑ:/ seemed to be the most difficult one since it showed the lowest success rate values in both cases (only a 8.9% in *Production*) and a 36.2% in *All productions* success rates in production and a 75.7% in discrimination). On the other hand, the phoneme /i:/ appeared to be the easiest one since it reached a 37.1% *All productions* and a 85.1% *Production* success rate values in production and a 95.1% success rate in discrimination. The number of times the learners requested the use of the TTS was also influenced by these differences: 648 for /ɑ:/ vs. 77 for /i:/.

Discrimination tasks							
	ɑ:	æ	ʌ	e	ɪ	i:	TPR (%)
ɑ:	***143***	34	12	-	-	-	75.7
æ	19	***198***	11	16	-	-	81.1
ʌ	20	14	***114***	-	-	-	77.0
e	-	11	-	***144***	12	-	86.2
ɪ	-	-	-	17	***105***	11	78.9
i:	-	-	-	-	4	***78***	95.1
PPV (%)	78.6	77.0	83.2	81.4	86.8	87.6	

Production tasks							
	ɑ:	æ	ʌ	e	ɪ	i:	TPR (%)
ɑ:	***151***	143	123	-	-	-	36.2
æ	78	***261***	35	113	-	-	53.6
ʌ	121	54	***127***	-	-	-	42.1
e	-	36	-	***195***	24	-	76.5
ɪ	-	-	-	33	***115***	49	58.4
i:	-	-	-	-	18	***103***	85.1
PPV (%)	43.1	52.8	44.6	57.2	73.2	67.8	

TABLE 7.16: Confusion matrices of the English Vowels prototype, adapted from [23]. Left table: confusion matrix of discrimination tasks (diagonal: right discrimination tasks). Right table: confusion matrix of production tasks (diagonal: right production tasks).

The CAPT system also revealed what the real difficulties of the users were in terms of the most difficult phonemes. The confusion matrices of discrimination and production tasks between the phonemes contrasted in each lesson are displayed in Table 7.16. In both tables the rows are the expected phonemes and the columns are the phonemes selected/produced by the user. *TPR* (true positive rate or recall) and *PPV* (positive predictive value or precision) are quality indicators. The symbol - stands for 'not applicable'.

In particular, the phoneme /ɑː/ was the hardest to predict in discrimination tasks (*TPR* = 75.7%) since it had the lowest recall; whereas /æ/ was the most commonly confused (*PPV* = 77.0%) since it had the lowest precision. The easiest phoneme in this type of tasks was /iː/ since it had the highest precision and recall (*PPV* = 87.6% and *TPR* = 95.1%). On the other hand, in production tasks, the phoneme /ɑː/ obtained the lowest precision, and also recall values (*PPV* = 43.1% and *TPR* = 36.2%); whereas /iː/ obtained the highest recall (*TPR* = 85.1%), and /ɪ/ the highest precision (*PPV* = 73.2%).

User's Behavior

The recommendation of specific training modes and activities was a part of the feedback in the training protocol which can also be analyzed. First, Table 7.17 represents the number of times that each training mode which affects the game score *G* was practiced (Discrimination, Pronunciation, and Mixed modes, see Section 7.4.10 for more details about this score). Three different scenarios can be described: (1) a training mode was passed (grade 60% or higher) at the first attempt, (2) a training mode was passed after repetition (because in previous attempts the user did not reach a 60% grade) following the recommended feedback or not, and (3) a training mode was not passed (with or without feedback).

	Proposed modes	Completed modes at first-attempt	Mode repetitions	Mode repetitions with feedback	Mode repetitions without feedback
Discrimination	60	51	10	6 [6,0]	4 [3,1]
Pronunciation	60	35	61	40 [23,17]	21 [2,19]
Mixed	60	43	34	12 [12,0]	22 [5,17]

TABLE 7.17: Comparison between following recommended feedback or not of the English Vowels prototype, adapted from [24]. Numbers between square brackets correspond to [passed, failed].

Clear differences between the three training modes are stated in Table 7.17. The Discrimination mode was the easiest one since it was passed 51 out of 60 times at the first attempt (83.33%). However, the Pronunciation training mode was the most difficult one, with 61 repetitions and a 58.33% success at the first attempt. When repeated, only in two occasions it was passed without the help of the provided feedback. Besides, the experiment showed significant differences ($U = 46.0$, $p < 0.001$, Mann–Whitney U test at 99% confidence level) between following or not the corrective feedback recommendations provided by the tool. In particular, these differences were higher in the case of the Pronunciation mode: without feedback, only a 10% of success was achieved; whereas the Mixed and Discrimination training modes achieved a 100% of success rate when the recommended feedback was followed.

Second, the effect of listening to a synthetic model of a misproduced word as an explicit corrective feedback in production events of Pronunciation and Mixed modes was also measured. Table 7.18 shows the number of production sequences which led to a positive or negative improvement. Each production sequence was given by: 1

	ɑː	æ	ʌ	e	ɪ	iː	
+	93 (32%) [2.8]	85 (32.4%) [3.2]	70 (35.4%) [2.9]	24 (30.4%) [3.1]	51 (47.2%) [2.9]	18 (58.1%) [4.2]	*341 (35.2%) [3.2]*
=	168 (57.7%) [0]	148 (56.5%) [0]	101 (51%) [0]	41 (51.9%) [0]	44 (40.7%) [0]	7 (22.6%) [0]	*509 (52.5%) [0]*
-	30 (10.3%) [-2.4]	29 (11.1%) [-2.7]	27 (13.6%) [-2.3]	14 (17.7%) [-2.9]	13 (12%) [-2.3]	6 (19.4%) [-1.7]	*119 (12.3%) [-2.4]*
	291 [0.1]	*262 [0.2]*	*198 [0.2]*	*79 [0.1]*	*108 [0.2]*	*31 [0.8]*	**969 [0.3]**

TABLE 7.18: Sequences of wrong production, listen, and repeat of the English Vowels prototype.

wrong production attempt followed by a number n ($n > 0$) of requested listenings, and ending in a new production attempt (from 1..4), always from the same word practiced. 969 sequences that comply with these requirements were registered of a total of 1779 (952+827, as showed in Table 7.15). There are three numbers in each cell of Table 7.18. The first one refers to the number of sequences. The second one is the percentage of sequences with respect to the phoneme (column). The last number is an indicator of positive (+), negative (-), or none of them (=) improvement in a scale of $[-4, 5]$ according to the target word's recognition position in the n–best list of hypotheses (1 to 5 positions, and 6 if not recognized). All right production events are included in the + row; whereas wrong production events can be included in any of three rows. When the subtraction of the n–best list position (in this case 5-best list) of the last recognition word attempt minus the first one is positive, the result is included in the + row, when it is negative in the - row, and when it is 0 in the = row.

When the last attempt was better than the first one, 3.2 points of difference were achieved. A slightly positive tendency of production improvement in the described sequences (0.3 out of 5) was confirmed as determined by a Wilcoxon signed-rank test at 95% confidence level since there were statistically significant differences between the (+) and (-) rows in Table 7.18 (Z = -10.362, $p < 0.001$). Results in Table 7.18 also corroborate results reported in Table 7.16 about the most difficult (/ɑː/) and easiest phonemes in the production activities (/iː, ɪ/).

Pre-test and Post-test Scores

As explained in Section 7.4.1, the pre-test and post-test content was identical (each student produced the same words in both tests), but the tests were performed at the beginning and at the end of the experiment. In order to assess the consistency of raters scores of both tests, a Kendall's coefficient analysis was carried out. A relevant inter-rater agreement was found (Kendall's coefficient W = 0.493; items = 900, raters = 3, p = 3.1e-19). A high correlation between the scores assigned by the raters to the speakers was also reported. In particular, the Pearson correlations between the mean scores assigned to the speakers in the pre-test are: $r = 0.87$, $p < 0.001$ between Rater1 and Rater2, $r = 0.73$, $p < 0.001$ between Rater1 and Rater3, and $r = 0.79$, $p < 0.001$ between Rater2 and Rater3. In the case of the post-test are: $r = 0.97$, $p < 0.001$ between Rater1 and Rater2, $r = 0.94$, $p < 0.001$ between Rater1 and Rater3, and $r = 0.95$, $p < 0.001$ between Rater2 and Rater3.

Table 7.19 shows the average scores assigned by the raters to the pre/post-test utterances. There were a total of 1200 scores for the in-classroom group (8 participants x 25 minimal pairs x 3 raters x 2 tests) and 1500 scores for the experimental group (10 participants x 25 minimal pairs x 3 raters x 2 tests). A comparison of pre-test and post-test scores, granted by the three human raters (column Rater: 1,2,3),

Group	Rater	Pre-test		Post-test		Difference (Wilcoxon signed-rank test)				
		mean	N	mean	N	mean	N	Z	r	p
Experimental	1	0.82	250	2.53	250	1.71	250	-7.864	0.50	<0.001
Experimental	2	0.99	250	2.45	250	1.46	250	-8.148	0.52	<0.001
Experimental	3	0.55	250	2.38	250	1.83	250	-7.422	0.47	<0.001
Experimental	1,2,3	0.85	750	2.59	750	1.74	750	-13.551	0.50	<0.001
In-classroom	1	0.41	200	0.68	200	0.27	200	-2.281	0.16	0.023
In-classroom	2	0.63	200	0.86	200	0.23	200	-3.056	0.22	0.002
In-classroom	3	0.27	200	0.61	200	0.34	200	-2.597	0.19	0.009
In-classroom	1,2,3	0.41	600	0.75	600	0.34	600	-4.566	0.20	<0.001

TABLE 7.19: Pre-test and post-test mean production scores of the English Vowels prototype, adapted from [23]. Mean is the average score assigned by a rater in a [0, 10] scale. The p-value is 2-tailed.

shows that there was improvement in both groups: from 0.85 to 2.59 in the experimental group, and from 0.41 to 0.75 in the in-classroom group. Since the content of pre-test and post-test was identical, a word–by–word comparison could be carried out between pre/post-test utterances of the same items by each student. In particular, a Wilcoxon signed-rank test found statistically significant differences between pronunciation improvement in both groups. The CAPT-group obtained an improvement of 1.74 points (Z = -13.551, $p < 0.001$), with a large effect size (r = 0.50); and the in-classroom group obtained a 0.34 improvement (Z = -4.566, $p < 0.001$) with a small effect size (r = 0.20).

In the case of the pre-test, results report that their scores were homogeneous at the beginning of the experiment since there were no statistically significant differences between groups scores (U = 18.0, p = 0.055, Mann–Whitney U test) with a moderate effect size (r = 0.46). However, there were significant differences between both groups in the scores of the post-test (U = 9.0, p = 0.001, Mann–Whitney U test), with a large effect size (r = 0.65). That means students who trained with the CAPT system achieved better pronunciation improvement values than learners in the in-classroom group regarding scores at the end of the experiment, both in absolute (1.74 vs. 0.34 of improvement) and in relative terms (205% vs. 82% of improvement).

Group	Post-test selected				Pre-test selected				Indifferent				NA
Experimental	73 (36.5%)				35 (17.5%)				79 (39.5%)				13 (6.5%)
	13%	47%	28%	9%	2%	48%	34%	14%	1%	2%	26%	79%	
In-classroom	25 (15.5%)				18 (11.3%)				109 (68.1%)				8 (5.0%)
	12%	28%	36%	24%	5%	27%	50%	16%	0%	2%	27%	73%	

TABLE 7.20: ABX test results of the English Vowels prototype. *Indifferent* indicates how often the rater shows no preference for any of the two stimuli. *NA* means 'no answer'. The second row of each group contains the Likert values (in percentage) assigned to each preferred option: from left to right, the left column means native like pronunciation and the right column means absolutely non-native pronunciation.

Table 7.20 shows the results of the ABX test (described in Section 7.4.10). Although six raters answered this test, only four of them showed a reasonable inter-rater reliability index (Fleiss Kappa index fair agreement, k = 0.393, Z = 13.8 , p

< 0.001) and were thus, kept to compute results in Table 7.20. A 60% of the experimental group samples received the highest marks in the post-test, against the 40% for the in-classroom group. A Pearson Chi–square test found statistically significant differences between the experimental and in-classroom group preferences ($\chi^2(2) = 30.461$, $p < 0.001$). Raters tended to prefer the post-test performances of the experimental group students: 73 vs. 35, ($p < 0.001$, Binomial test), contrasting with the in-classroom group results: 25 vs. 18 (with no statistically significant differences).

ID	Rater1 score	Rater2 score	Rater3 score	Raters score (mean)	Game score (G)
1	5.52	5.94	5.92	5.79	8.30
2	3.76	3.92	4.52	4.07	7.80
3	3.62	3.90	2.54	3.35	8.10
4	2.40	2.62	2.56	2.53	8.10
5	2.24	2.54	2.52	2.43	7.40
6	2.16	3.04	2.32	2.51	7.80
7	2.12	1.58	0.90	1.53	7.40
8	2.08	2.28	2.04	2.13	7.40
9	1.28	1.96	0.32	1.19	7.30
10	0.44	0.70	0.18	0.44	7.10

TABLE 7.21: Correlation between the software and human raters post-test scores of the English Vowels prototype, adapted from [23]. Learners belong to the experimental group. Score's scale is [0, 10]. ID is the user identifier.

Subjective scores coming from the pre- and post-test correction by experts and objective ones were compared in Table 7.21. The average scores assigned by each rater to the subjects and the game score obtained by each one of them training with the CAPT system are displayed. Individually, the Pearson correlation between the game score and each rater was $r = 0.84$ for Rater1 ($p = 0.002$), $r = 0.86$ for Rater2 ($p = 0.001$) and $r = 0.79$ for Rater3 ($p = 0.007$). Besides, the correlation of Rater1, Rater2, and Rater3 together was $r = 0.84$ ($p = 0.002$). Finally, a potential rater score, Rater', can be obtained from the CAPT tool score (G), with an average error of $\pm 5.5\%$ using a linear regression model (see Figure 7.18):

$$Rater' = -21.724 + 3.171 * G \tag{7.9}$$

In order to further validate the results about improvement between pre- and post-tests, a correlation study was carried out between the time spent by students to fulfill the test and the results of this test. Each participant took an average of 79.72 seconds to complete the pre-test (59 seconds min. and 107 seconds max.) and an average of 95.61 seconds to complete the post-test (62 and 140 seconds min. and max.). A moderate correlation was found ($r = 0.506$, $p = 0.032$; and $r = 0.459$, $p = 0.055$, respectively) between the time that students spent on the post-test, on the one hand, and their performance (human raters score in the post-test) and achieved learning (difference between post-test and pre-test human raters score), on the other. These values suggest a certain impact of the time spent on the post-test over user's performance and learning, but although post-test time was generally higher than pre-test time, the correlation between pre-test and post-test time, calculated as $r = 0.75$, $p < 0.001$, also suggests a dependence on the speaker. These results suggest that a proper

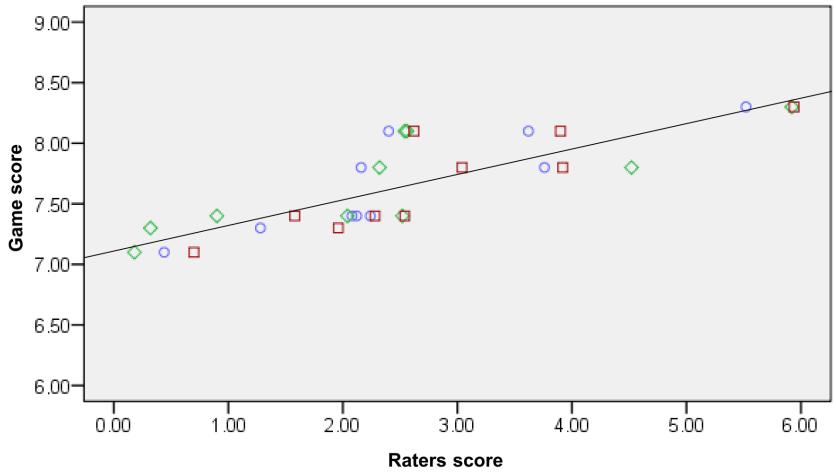

FIGURE 7.18: Correlation between the game and human raters post-test scores of the English Vowels prototype, adapted from [23]. Circles, squares, and rhombuses represent the Rater1, Rater2, and Rater3 average scores, respectively.

evaluation of the incidence of test completion time requires further and rigorous experimentation to get undeniable conclusions.

7.4.13 Japañol Prototype Results

Results extracted from this prototype tried to reinforce the same answers to the research questions of the previous prototype, *English Vowels*: RQ2 and RQ1 (and their *Issues*), following the steps defined in the research objectives RO1, RO2, RO3, and RO4.

Results related to participants from the Language Center of the University of Valladolid have been partially published in [25], [26], [27]. Other contributions related to this prototype have been partially published in [28], [29].

User's Performance

Results related to the interaction with the CAPT system by users of the experimental group (18) are displayed in Table 7.22. $\overline{n}$, m, and M are the mean, minimum and maximum values, respectively. *Time (min)* row stands for the time spent on minutes per learner in each mode in the three sessions of the experiment. *#Tries* represents the number of times a mode was executed by each user. The symbol - stands for 'not applicable'. *Mand.* and *Req.* mean mandatory and requested (listening). The TTS system was employed in both listening types; whereas the ASR was only used in the *#Productions* row.

Japañol learners spent an average of 100.56 minutes performing the proposed activities in the three training sessions. A 85.25% of this time was consumed by carrying out interactive training modes (Exposure, Discrimination, Pronunciation, and Mixed modes). As a mean term, users listened to the TTS system 612.6 times and produced 291.8 times with the ASR system, reaching a rate of 9.0 uses of the TTS/ASR per minute.

Important differences in the level of use of the tool depending on the user are also illustrated in Table 7.22. For instance, the fastest learner performing pronunciation

	Theory			Exposure			Discrimination			Pronunciation			Mixed		
	$\overline{n}$	m	M	$\overline{n}$	m	M	$\overline{n}$	m	M	$\overline{n}$	m	M	$\overline{n}$	m	M
Time (min)	14.80	8.7	20.8	19.7	12.8	21.9	7.1	4.1	13.8	42.6	22.4	72.9	16.4	7.6	30.5
#Tries	7.8	6	10	10.6	7	16	8.5	7	15	9.8	7	14	6.7	3	10
#Mand.List.	-	-	-	287.8	210	390	91.7	70	134	-	-	-	20.2	9	30
#Req.List.	-	-	-	99.3	53	157	33.0	0	153	54.9	0	127	25.6	6	60
#Discriminations	-	-	-	-	-	-	91.7	70	134	-	-	-	20.2	9	30
#Productions	-	-	-	-	-	-	-	-	-	208.8	116	356	82.9	38	181
#Recordings	-	-	-	62.4	42	81	-	-	-	-	-	-	-	-	-

TABLE 7.22: User's performance with the CAPT system of the Japañol prototype.

activities spent 22.43 minutes; whereas the slowest one took 72.85 minutes. This contrast can also be observed in the time spent on the rest of the training modes and in the number of times learners practice each one of them (row *#Tries*). The inter-user differences affect both the number of times the users made use of the ASR (154 minimum vs. 537 maximum) and the number of times they requested the use of TTS (59 vs. 497 times).

Task	Event	[fl]	[fɾ]	[l]	[ɾ]	[rr]	[s]	[θ]	[f]	[fu]	[xu]	Total
Disc.	Right t-s	123 (65.8%)	115 (62.5%)	239 (76.1%)	217 (71.4%)	215 (85.7%)	95 (74.8%)	214 (89.2%)	104 (96.3%)	115 (77.2%)	111 (74.0%)	1548 (76.9%)
Disc.	Wrong t-s	64 (34.2%)	69 (37.5%)	75 (23.9%)	87 (28.6%)	36 (14.3%)	32 (25.2%)	26 (10.8%)	4 (3.7%)	34 (22.8%)	39 (26.0%)	466 (23.1%)
Disc.	Mand.List.	187	184	314	304	251	127	240	108	149	150	2014
Disc.	Req.List.	65	52	139	115	51	45	45	16	89	103	720
Prod.	Right t-e	170 (31.8%)	137 (25.1%)	253 (45.1%)	289 (52.6%)	252 (51.5%)	134 (24.6%)	226 (28.4%)	116 (51.8%)	138 (19.2%)	140 (21.4%)	1855 (33.0%)
Prod.	Wrong t-e	364 (68.2%)	408 (74.9%)	308 (54.9%)	260 (47.4%)	237 (48.5%)	410 (75.4%)	571 (71.6%)	108 (48.2%)	580 (80.8%)	513 (78.6%)	3759 (67.0%)
Prod.	Right t-s	170 (78.3%)	137 (68.2%)	253 (84.9%)	289 (89.2%)	252 (89.0%)	134 (66.7%)	226 (67.7%)	116 (85.9%)	138 (56.6%)	140 (60.6%)	1855 (75.2%)
Prod.	Wrong t-s	47 (21.7%)	64 (31.8%)	45 (15.1%)	35 (10.8%)	31 (11.0%)	67 (33.3%)	108 (32.3%)	19 (14.1%)	106 (43.4%)	91 (39.4%)	613 (24.8%)
Prod.	Mand.List.	-	-	-	-	-	-	-	-	-	-	-
Prod.	Req.List.	128	125	105	103	70	146	202	29	240	186	1334

TABLE 7.23: Right, wrong, and listening events categorized by sounds of the Japañol prototype.

In Table 7.23 Disc. and Prod. correspond to discrimination and production task-types, respectively. *Right t-s* refers to a correct task sequence. *Wrong t-s* means incorrect task sequence: in Disc. it refers to a wrong attempt of discrimination; in Prod. it refers to five misproduction task-events. *Right/Wrong t-e* are correct/incorrect single production events. A *wrong t-e* occurs when the ASR does not include the produced word (or a homophone) in the first position of the *n*–best list of hypotheses. *Mand.* and *Req.* mean mandatory and requested (listening). The symbol - stands for 'not applicable'.

Analyzing all perception and production events registered in Discrimination, Pronunciation, and Mixed modes it was evidenced that although their sequence success rate values, *Right t-s*, were similar (76.9% and 75.2%, respectively), the difference increased notably when compared to the single events rate, *Right t-e* (76.9% vs. 33.0%, discrimination and production, respectively). As a reader's reminder, while a discrimination task conveyed a single attempt, up to five attempts could be conducted before a production task was passed.

Table 7.23 also reveals which sounds were the most difficult ones for the learners since there were important differences that affect both discrimination and production activities. For instance, discrimination success rates related to words with sounds [fl] and [fɾ] reached the lowest values (65.8% and 62.5%, respectively); whereas sound [f] (when contrasting to [θ]) seemed to be the easiest one (96.3% success rate). In the case of production activities, sounds [fu] and [xu] showed the lowest production success rates values, both in sequences (56.6% and 60.6%, respectively) and single events (19.2% and 21.4%, respectively). The number of times the TTS was

requested for these sounds was also very high (240 and 186 times, respectively). However, users barely found problems with sounds [ɾ] and [rr], both in sequences (89.2% and 89.0%, respectively) and single events (52.6% and 51.5%, respectively); being the TTS rarely requested (103 and 70 times, respectively).

On the other hand, sounds [rr] and [f] (when contrasting to [θ]) appeared to be the easiest ones in general, since they reached a 51.5% and 51.8% *Prod. Right t-e* and a 89.0% and 85.9% *Prod. Right t-s* success rate values in production and a 85.7% and 96.3% success rate in discrimination tasks. The low number of times the use of the TTS was requested for these two sounds reinforced the idea that users found easy to perceive and produce them (70 and 29 times, respectively).

Discrimination tasks												
#Lis		[fl]	[fɾ]	[l]	[ɾ]	[rr]	[s]	[θ]	[f]	[fu]	[xu]	TPR (%)
65	[fl]	***123***	64	-	-	-	-	-	-	-	-	65.8%
52	[fɾ]	69	***115***	-	-	-	-	-	-	-	-	62.5%
139	[l]	-	-	***239***	56	19	-	-	-	-	-	76.1%
115	[ɾ]	-	-	71	***217***	16	-	-	-	-	-	71.4%
51	[rr]	-	-	15	21	***215***	-	-	-	-	-	85.7%
45	[s]	-	-	-	-	-	***95***	32	-	-	-	74.8%
45	[θ]	-	-	-	-	-	15	***214***	11	-	-	89.2%
16	[f]	-	-	-	-	-	-	4	***104***	-	-	96.3%
89	[fu]	-	-	-	-	-	-	-	-	***115***	34	77.2%
103	[xu]	-	-	-	-	-	-	-	-	39	***111***	74.0%

TABLE 7.24: Confusion matrix of discrimination tasks of the Japañol prototype (diagonal: right discrimination tasks).

Table 7.23 considers the sounds individually. Another approximation is to compare the sounds of each lesson since they were presented in pairs in the CAPT system. Tables 7.24 and 7.25 show the confusion matrices between the sounds of the minimal pairs in perception and production events. In both tables the rows are the phonemes expected by the tool and the columns are the phonemes selected (discrimination) or produced (production) by the user. *TPR* is the true positive rate or recall. The symbol - stands for 'not applicable'. *#Lis* is the number of requested listenings to the sound row.

In agreement with the results presented in Table 7.23, the most confused pairs in discrimination tasks were [l]–[ɾ], both individually (127 times) and preceded by the sound [f] (132 times). Besides, the number of requested listenings related to these sounds was the highest one (204 times for [l] and 167 for [ɾ]). The least confused pair in discrimination was [θ]–[f] (15 times). The sounds with the lowest discrimination *TPR* rate were [fl] and [fɾ] (both $< 66.0\%$), and those with the highest discrimination *TPR* rate were [θ] and [f] (both $> 89\%$), corresponding to the lowest number of requested listenings (45 and 16, respectively).

Table 7.25 shows the results related to production events per word utterance. A positive improvement from first to last attempt was observed (final column), being the highest ones [fl] (33.2%) and [fɾ] (21.1%) sounds. In particular, these two sounds constituted the most confused pair in first attempt production tasks (152 times), where the least confused one was [l]–[rr] (59 times). The sounds with the lowest production *TPR* rate were [fl] and [s] (both $< 46\%$), and those with the highest production *TPR* rates were [ɾ] and [rr] (both $> 73\%$).

Production tasks (first attempt | last attempt)												
#Lis		[fl]	[fɾ]	[l]	[ɾ]	[rr]	[s]	[θ]	[f]	[fu]	[xu]	TPR (%)
13|128	[fl]	***65***|***170***	79|47	-	-	-	-	-	-	-	-	45.1%|78.3%
3|125	[fɾ]	73|64	***65***|***137***	-	-	-	-	-	-	-	-	47.1%|68.2%
9|105	[l]	-	-	***177***|***253***	45|31	37|14	-	-	-	-	-	68.3%|84.9%
8|103	[ɾ]	-	-	33|21	***209***|***289***	42|14	-	-	-	-	-	73.6%|89.2%
3|70	[rr]	-	-	22|9	44|22	***189***|***252***	-	-	-	-	-	74.1%|89.0%
6|146	[s]	-	-	-	-	-	***58***|***134***	66|67	-	-	-	46.8%|66.7%
2|202	[θ]	-	-	-	-	-	79|96	***142***|***226***	38|12	-	-	54.8%|67.7%
0|29	[f]	-	-	-	-	-	-	38|19	***97***|***116***	-	-	71.9%|85.9%
4|240	[fu]	-	-	-	-	-	-	-	-	***62***|***138***	62|106	50.0%|56.6%
5|186	[xu]	-	-	-	-	-	-	-	-	59|91	***63***|***140***	51.6%|60.6%

TABLE 7.25: Confusion matrix of production tasks of the Japañol prototype at first and last attempt per word sequence (diagonal: right production tasks at first and last attempt per word sequence). *#Lis* is the number of requested listenings of the corresponding *sound* row at (*first* | *last*) attempt.

On the other hand, the most confused pair in last attempt production tasks was [fu]–[xu] (197 times), reaching the lowest production *TPR* rates (56.6% and 60.6%, respectively). Besides, the number of requested listenings was the highest one in both cases (240 and 186, respectively). The least confused pair was [l]–[rr] (14 times), reaching *TPR* rate values higher than 85%.

Pre-test and Post-test Scores

As explained in Section 7.4.1, the pre-test and post-test content was the same (each student produced the same words in both tests), but the tests were performed at the beginning and at the end of the experiment. 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.). These trends agreed with those reported for the English Vowels prototype (see the last paragraph of Section 7.4.12).

As a **first** approach, a perceptual test of the utterances from the pre-test and post-test with the 8 participants of **Language Center of Valladolid** was carried out. The scores assigned by the five raters were correlated with those obtained in the CAPT system (Game score, *G*). There were a total of 896 utterances (8 participants x 28 minimal pairs x 2 words per minimal pair x 2 tests). Table 7.26 shows the scores for each user at any of the given stages of the experiment (pre-test, CAPT tool, post-test, and a delta score of the pre and post-test). ASR and RATER scores refer to the learners' qualifications of pre/post tests by Google ASR and the average value of the human raters average scores, respectively (see Section 7.4.10 for scoring details).

Concerning to raters scores (RATER), a consistency test among them based on the Fleiss' Kappa statistical indicator was carried out both for pre-test (substantial agreement, $k = 0.63$) and post-test (moderate agreement, $k = 0.50$) evaluations. Comparing subjective and objective scores, delta score values (the last two columns) were positive in almost all cases both in RATER and ASR (only two of them decrease in RATER, both in top three). They also showed a fair correlation with pre-test expert scoring (r = -0.856) and post-test expert scores with ASR ones (r = -0.735). Pre-test scores assigned by experts (RATER) showed a reasonable linear regression correlation with those obtained by applying the ASR in the same test (r = 0.890). A similar

	Pre-test		Game score	Post-test		Δ (Post - Pre)	
ID	RATER	ASR	*G*	RATER	ASR	RATER	ASR
07	9.8	5.1	9.4	9.4	7.5	-0.4	2.4
05	8.6	4.5	9.1	8.8	4.6	0.2	0.2
01	8.1	2.9	8.0	7.9	4.6	-0.3	1.7
02	7.3	3.7	8.4	7.8	4.8	0.4	1.2
03	6.9	1.9	6.3	7.6	2.7	0.8	0.8
08	6.8	2.8	6.5	7.1	3.4	0.4	0.6
06	5.8	0.8	5.9	6.5	3.2	0.8	2.4
04	5.4	2.1	6.9	7.3	2.9	1.9	0.8

TABLE 7.26: Scores at different stages of the Japañol prototype, adapted from [25]. The columns indicate the rater (experts, ASR system or game). Score's scale is [0, 10].

correlation was found for post-test RATER and ASR (r = 0.834). Game scores (G) showed good correlation with RATER post-test results (r = 0.912), clearly over the correlation found between G and pre-test human rater results (r = 0.867).

Second, we analyzed the pre- and post-test utterances of the 25 participants of **University of Seisen** with Kaldi and Google ASR systems, since we wanted to find pronunciation mistakes associated with key features of proficiency level characterization (with Kaldi), and, in this case, we did not have access to enough human resources to carry out the perceptual assessment of the audio files. The specific-purpose Kaldi system allowed us to assess utterances automatically without depending on the general-purpose Google ASR system which works as a black-box. To achieve this goal, we recorded a dataset of 41,000 utterances in total with 10 native Spanish speakers (5 women and 5 men) who read 25 repetitions of 164 words of the minimal pairs related with the experiment (see details of the dataset in Table D.1 in Appendix D).

In order to design our custom ASR system with Kaldi (see Section E.2 for details about the elaboration), six different phoneme-level train models were tested with the audio dataset recorded with native speakers before assessing the non-native test utterances. The *All* model included 41,000 utterances of the native speakers in the train set. The *Women* model included 20,500 utterances of the 5 female native speakers in the train set. The *Men* model included 20,500 utterances of the 5 male native speakers in the train set. The *BestNative1*, *BestNative2*, and *BestNative3* 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 word error rate (WER) values of all possible 80%/20% combinations (training/test) of the native speakers (i.e., 4 female and 4 male native speakers for training: 8, and 1 female and 1 male for testing: 2), and choosing the best three WER values. Non-native test model consisted of 2800 utterances (25 participants x 28 minimal pairs x 2 words per minimal pair x 2 tests). Table 7.27 displays the WER values for the 4-gram Kaldi triphone (tri4) models described above (see more details about triphones in Section E.2). The *All* model reported the best test results. However, we chose the *Women* train set for elaborating our ASR system with Kaldi since the test utterances of the University of Seisen were all female speakers.

Table 7.28 shows the average scores assigned by the Google (GCSTT) and Kaldi (*Women* train model) ASR systems to the 2,800 utterances of the pre/post-tests (1,400 + 1,400) classified by the three groups of participants, in a [0, 10] scale. The learners

	Test model	
Train model	Native	Non-native
All	0.0024	44.22
Women	3.10	55.91
Men	1.55	64.12
BestNative1	0.14	46.40
BestNative2	0.14	46.98
BestNative3	0.23	48.08

TABLE 7.27: WER values (%) of the six models tested for the Kaldi ASR system.

who trained with the Japañol tool (experimental group) achieved the best pronunciation improvement values in both Google (0.7) and Kaldi (1.1) systems. However, the in-classroom group achieved better results in both tests and in 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) and pronunciation improvement values (0.2 and 0.5).

	Pre-test				Post-test				Δ (Post - Pre) – Wilcoxon signed-rank test					
	Google		**Kaldi**		**Google**		**Kaldi**		**Google**			**Kaldi**		
Group	*G*	N	*G*	N	*G*	N	*G*	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	3.0	392	3.2	392	3.5	392	0.2	0.002	-3.154	0.5	0.059	-1.891

TABLE 7.28: Google and Kaldi results of the tests utterances of the Japañol prototype. *G*, *N*, and Δ refer to Game score, number of utterances, and difference, 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 values. In the case of the placebo group, there were differences only in the Google ASR values (see *p* and *Z* values in Table 7.28). 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 ($p < 0.001$; $Z = -5.324$). Post-test differences between the in-classroom and placebo groups were only found in the Kaldi scores ($p < 0.001$; $Z = -7.651$). Although there were statistically 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 are minimal since the effect size values were small ($r = 0.10$ and $r = 0.20$, respectively).

Finally, we analyzed several correlations between the pre/post-test scores of both ASR systems (all groups) and the *G* score with the post-test scores of both systems (only with the experimental group) in order to compare the three sources of objective scoring. First, Figure 7.19 represents the moderate positive Pearson correlations found between the Google and Kaldi post-test ($r = 0.57$, $p = 0.002$) and pre-test ($r = 0.51$, $p = 0.005$) scores.

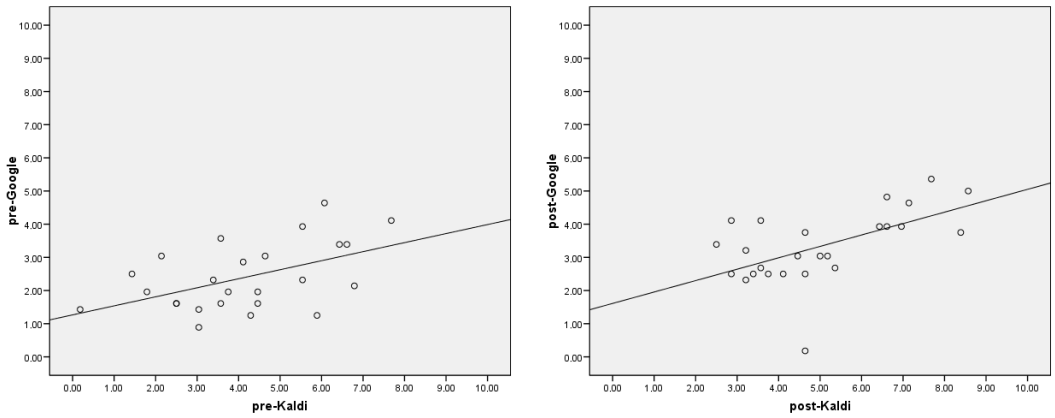

FIGURE 7.19: Correlation between the Google and Kaldi ASR scores of the pre-test (left) and post-test (right) of the Japañol prototype.

Second, Figure 7.20 represents the fairly strong positive Pearson correlations found between the *G* scores and the Google ($r = 0.81$, $p = 0.002$) and Kaldi ($r = 0.74$, $p = 0.007$) ASR systems post-test scores.

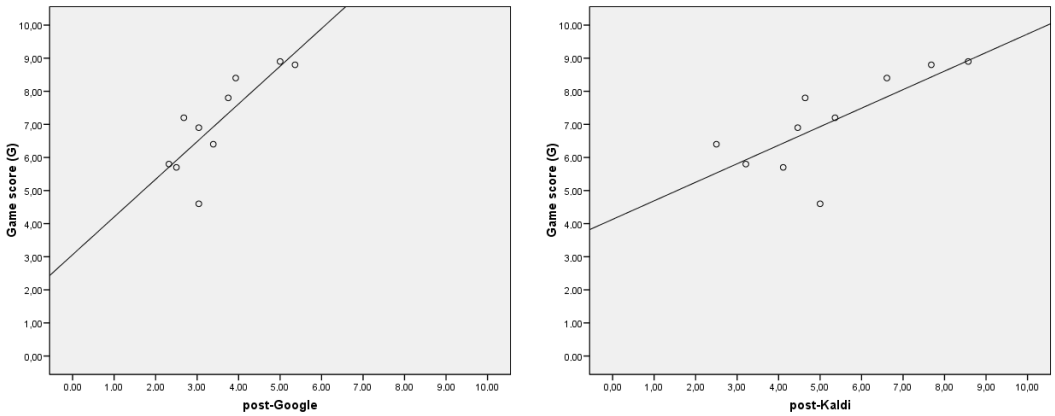

FIGURE 7.20: Correlation between the Google (left) and Kaldi (right) ASR scores of the post-test with the game score of the Japañol prototype.

7.5 Competitive Learning Experiment

The training approach in the previous experiments was individualistic in all cases. Participants performed different activities with a CAPT tool without sharing tasks with other participants. At this point we considered the possibility of including challenges between participants in order to help to improve pronunciation, motivation, and performance. Thus, we decided to carry out one last experiment, *Competitive Learning*, in which a second version of the TipTopTalk! CAPT tool was designed as a result of an evolved process, named COP. The novelty was the possibility of learners to "challenge" others in a social competition, since allowing users to play alone and letting them to choose the activities they wanted to practice in TipTopTalk! led us to detect a stagnation in training intensity and a pronunciation decrease for the most active users during the last days of the competition (as explained in Figure 7.10).

In this new version of the game, learners could challenge each other by performing some pronunciation activities in the CAPT system with current general-purpose speech technology, in order to climb up a leaderboard. Native Spanish students from the University of Valladolid participated in this EFL pronunciation competition. The great quantity of data gathered was intended not only for analyzing the effects of the explicit competition on user's motivation, performance, and learning, but also for collecting a significant number of spoken data for future development (see Section 9.2). Participants' results were statistically analyzed according to their motivation, performance, and learning outcomes in the competition.

7.5.1 Experimental Procedure

A one-month protocol was followed in this experiment. It included a pre-quest, a competition period, a post-quest, and four focus group sessions, as shown in Figure 7.21. Firstly, the recruitment campaign was active during six days. At the same time users registered into the experiment, they had to complete the pre-quest via a web. Once the recruitment campaign was over, the competition began for all enrolled learners (see Section 7.5.3 for more details about participants) and lasted for 24 days. Subjects could take part in the competition anytime anywhere with their own devices during the protocol's interval dates. Results obtained during the competition were gathered into log files. Audio data was also kept for future analysis. When the competition was over, users are required to complete a post-quest (also via web) until a maximum time of 4 days. However, users who did not complete all required stages had to fill in a different post-quest from the rest of participants about the reasons for abandoning.

Furthermore, one week after the end of the competition four one-hour focus group sessions with 16 participants in each one of them were conducted. This session was carried out at the university facilities, with all participants sitting around a table face to face. One member of the research group conducted the session while other took notes and recorded the audio of the session. Firstly, an overview about the results obtained during the experiment was presented to the participants during 10 minutes. Then, subjects were asked about their perceptions and opinions about the experiment with the possibility of discussion with other participants (30 minutes). The last 20 minutes of the session were intended to suggest improvements and future work.

7.5.2 Enrollment

Students were asked to participate voluntarily in this experiment, via invitation emails to their corporate university email address and by means of invitation talks in selected classrooms. They registered in the competition by filling in a registration form with some personal information and signing an informed consent. Additionally, they had to complete some pre/post-competition questionnaires at the beginning and at the end of the competition (see Section 7.5.5), respectively. Some of the students were randomly selected at the end of the experiment to participate mandatorily on a focus group session. After registering, students received the instructions to download the software prototype from Google Play. A starting and ending date for playing were established for a total of 24 days of competition.

Different kinds of reward were offered to participants depending on their level of participation. In particular, a diploma and an academic certification were given

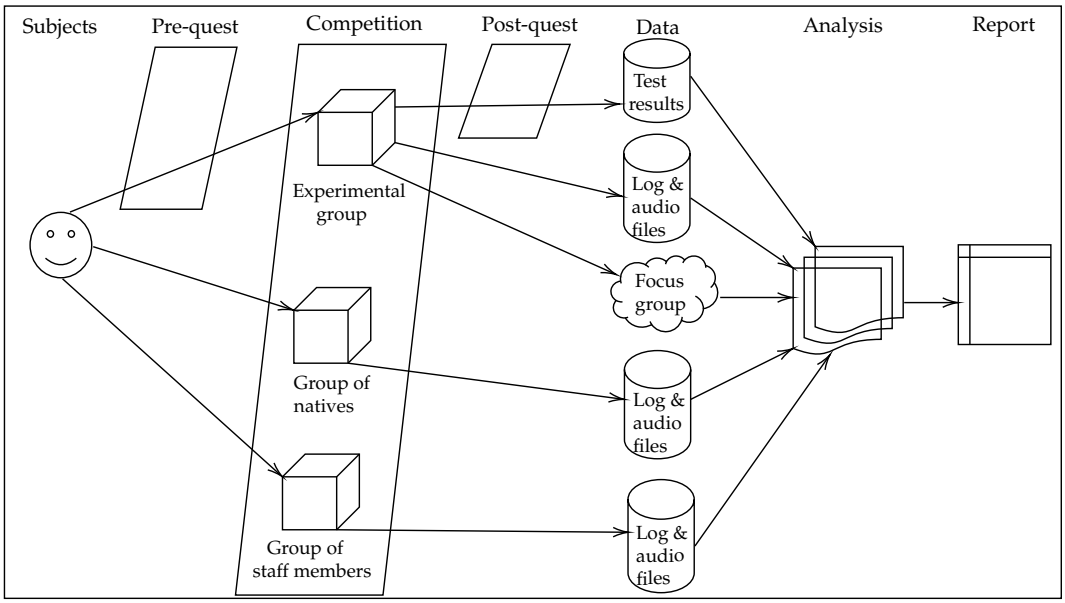

FIGURE 7.21: Steps of the COP prototype's protocol.

to participants who completed at least 60 challenges. A prize was also awarded to the first 15 classified (the better the position of the leaderboard the higher the prize). Finally, during the whole competition, the research team was available to answer emails from users asking for technical help about the installation and execution of the games.

7.5.3 Participants

Initially, **354** users signed up online, completed the pre-quest, and started to compete; being **165** the final number of users who performed every stage of the experiment and neither belonged to the research staff (**5**) nor were native (**2**): registration, pre-quest, competition, and post-quest stages. There were some subjects who took part in this experiment as support (see Table B.7 for specific details). The rest of participants were native Spanish students at University level. They studied EFL for several years during primary and secondary school. In summary, subjects related to this experiment can be classified as follows:

1. **Native Spanish speakers who fully participated** (165). They participated during the whole competition. They were awarded with a diploma and an academic certification if they completed 60 or more challenges. The best 15 users of the leaderboard received a reward. 64 speakers took part in the four focus group sessions (16 learners in each one). They were 111 women and 54 men (average age = 21.44, *SD* = 1.82).

2. **Native Spanish speakers who abandoned before the end** (182). They were subjects who completed the registration form and the pre-quest, participated some days in the competition without achieving the academic certification, and completed the post-quest about reasons of early abandonment. They were 100 women and 82 men (average age = 21.01, *SD* = 1.42).

3. **Natives** (2). They participated only during the first seven days of the competition. They were intended to serve as bait for the rest of the users. They were

awarded with a reward. They were 1 woman and 1 man from USA (average age = 22.05, *SD* = 0.5).

4. **Staff members** (5). They took part only during the first seven days of the competition. They were intended to motivate users by submitting and accepting a great quantity of challenges to avoid a possible initial stagnation of the dynamics of the competition. They were 5 men from the ECA-SIMM group.

7.5.4 COP CAPT System Description

The competition was run via a gamified CAPT tool for smart devices, called Clash of Pronunciations, COP. This new version of the CAPT system presented in the Non-guided Learning experiment (see Section 7.3.4) consisted in a turn-based social game in which users challenged each other and their results were reflected on a leaderboard. The main goal of a user playing with the developed CAPT tool was to achieve points by performing some pronunciation **activities** based on the NCM (see Section 2.1.1) in **matches** of **challenges**, trying to reach the best position possible in a **leaderboard**. A match could be played in two modes: **Playing** and **Training**. In the Playing mode, users got points by participating in challenges against other subjects (main difference with the Non-guided Learning experiment's CAPT system). Each challenge involved a minimum of two and a maximum of five participants who performed the same activities in their respective matches. They also included twelve discrimination and production activities, in rows of two. In the Training mode, users played matches individually, which included exposure, discrimination or production activities. However, in this mode users did not get rewards nor points. In this experiment was included the Google ASR, GCSTT, and TTS systems.

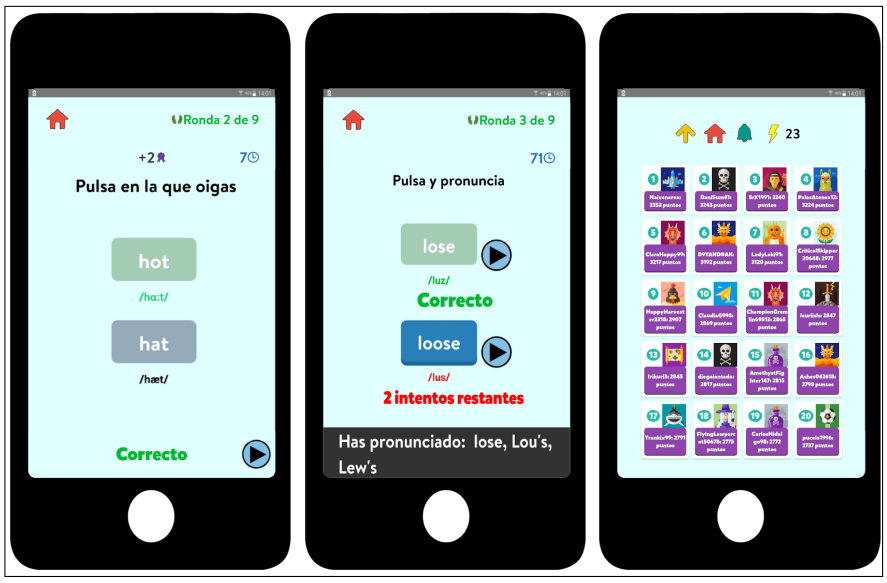

FIGURE 7.22: COP CAPT system screenshots of discrimination (first picture) and production (second picture) activities in a match. Leaderboard example (third picture). Adapted from [31].

The CAPT tool was populated with a database of 329 American English minimal pairs of vowel and consonant contrasts (English words and their phonetic transcription)[6]. These words were organized into lists of ten or more pairs of words, which

[6] https://github.com/eca-simm/minimal-pairs-cop

each one of them corresponded to a pair of phonemes to contrast. The minimal pairs list to be used in the activities of a match was randomly selected by the system to try to keep the same variety of the difficulty level along the competition. In the discrimination activities of the Playing mode (see the first screenshot of Figure 7.22), users could listen to the sound as many times as they want, although the final score was penalized after the second listening. In the production activities of the Playing mode (see the second screenshot of Figure 7.22), there was a maximum number of three pronunciation attempts per word of the pair with a penalization after the second attempt. Additionally, the system invited users to listen to the correct pronunciation without penalization in the scoring. A production attempt was considered correct (right) when the orthographic transcription of the word (or some homophone) was included in the three first positions of the text hypotheses of the ASR result. There was a time limit of 100 and 10 seconds per production and discrimination activity, respectively. In the Training mode, in both games, users could freely select the list of minimal pairs to train and the activity type (exposure, discrimination, or production). In particular, there was no time limit and users cannot obtain points. Finally, users were rewarded with digital trophies (badges) and motivated with inspirational push messages sent to the CAPT system from the web server to keep users playing and training.

In order to establish a competitive game configuration, the guidelines stated by [123] were followed (see Section 5.1 for more details about competitive scenarios in learning). That is, a player to participate in the Playing mode must create a new challenge or accept the invitation of a challenge sent by another users, trying to beat them (*interaction with other parties*). A challenge starts with the match issued by the player who creates the challenge against a set of selected users. All subjects of a given challenge perform the same discrimination and production activities included in the match in their respective match (nine activities, six for discrimination and three for production, interspersed). In each activity, the points obtained by the users depend directly on the quality of their performance (two points per first-time right attempt and one point in other right attempt. There are no negative points). Players also receive extra points when they beat players with a higher position on the leaderboard. These points are valid when the challenge is finished, that is, when the last user of the challenge performed her/his match. Then, the winner(s) and loser(s) of the challenge (*negative goal interdependence*) are declared; updating the leaderboard of the competition with their final scores (*comparability among participants*). See an example of the leaderboard in the last screenshot of Figure 7.22). The winner of the competition is the player who achieves more points during all the competition days, reaching the first position of the leaderboard. The winner of a challenge is the player who achieves the highest MatchScore. Ties can occur in challenges involving more than two players (*winning rules*). In order to guarantee a similar game level, the possible available opponents for a challenge belong to a range of ten positions above and below the creator's current leaderboard position. There is a limit of 30 matches per user per day in order to avoid counterproductive extra working load. Only players who complete at least 60 challenges obtain an academic certification (*perceived scarcity*). Also, subjects who reach one of the fifteen first position on the leaderboard at the end of the competition obtain a reward (*quantity of winners*).

To summarize, a player in this competition has the following options:

- **Submit (create) challenges**. Each user can challenge up to four other learners (from those who are 10 positions below or above her/him of the leaderboard).

The user who initiates the challenge plays its first match and wait for the answer of the rest of users.

- **Play matches in challenges**. Users must perform different activities with minimal pairs and obtain a score. When a player finishes the match, a message with the points and right/wrong attempts is displayed. Then, the player must wait for the other participants' results in order to declare winners and losers of the challenge and add the points achieved to the leaderboard.
- **Train**. In addition to playing matches, users can choose training activities as an unlimited option. They can choose exposure, discrimination, or production activities of the minimal pair contrasts they want. They do not add points to the leaderboard in this mode.
- **Respond to the received challenges**. The user receives a message about the playmate who is challenging her/him. The incoming challenge can be accepted to perform the match or ignored without counting into the restriction of a maximum of 30 challenges per day (although one is deducted from the creator of the challenge).

7.5.5 Instruments

Different sources of data for this experiment can be discerned:

- **Registration forms**: user's demographic information, such as name, age, gender, L1, academic level, and final consent to analyze all gathered data. This information was carefully collected and saved into digital text documents.
- **Pre-quest:** an online questionnaire about three different categories of questions. The first category included Likert-scale type questions to evaluate the degree of competitiveness of the users. Second, the scale of scholar motivation for (EME-E), subdivided into extrinsic and intrinsic motivation adapted and validated in Spain [197]. Finally, a specific questionnaire for evaluating the self-concept in what concerns to pronunciation and discrimination of sounds level in the context of SLA [198]. This data was gathered into a secure web server.
- **User's interaction log files**. The CAPT tool gathered data associated with all low-level interaction events and monitors all user activities. This data was saved into local log files and automatically uploaded to a web server (see Section 7.5.6 for more details).
- **Audio recordings**. In this experiment the use of the basic–free Google ASR system for Android and the online speech API called GCSTT was alternated. The latter system allowed us to keep the audio files sent to their speech recognition system. This data was gathered into a secure web server.
- **Post-quest**: an online questionnaire for users who finished the corresponding competition about the usability of the tool (adapted from [199]). Three different questionnaires about reasons for playing, attitude toward competition, and information from users who abandoned the game before completing all the stages were also included. This data was collected and saved into a secure web server.

- **Focus group sessions**. The audio of the session was recorded via a camera and the most important quotes and requests of the participants were written by a member of the research team by taking notes. This data was carefully collected and saved into digital text documents.

7.5.6 Metrics

Game Intensity and Motivation

User's game intensity was characterized in terms of declared reasons for participating in the competition (motivation), and quantity and regularity in matches' participation. Data related to user's motivation was gathered in log files:

- **Number of active days**: amount of days in which a user participates in Training or Playing matches.
- **Number of attempts**: amount of discrimination and production activities performed by a user in Training or Playing matches.
- **Degree of motivation**: subjective answers to the questionnaire at the end of the competition (motivation for participating, feelings during the competition, and reasons for abandonment).

Performance

Related to the amount of events tracked from each participant. Different indicators of the CAPT tool characterize user's performance:

- **Production attempt**: every attempt of producing correctly the proposed word of a pair. Binary value (true, false) indicating whether the orthographic transcription of the word matches to the user's utterance result of the *n*–best list of hypotheses of the ASR.
- **Production success rate**: percentage of right production attempts according to the total number of attempts of the user in the competition.
- **Discrimination attempt**: every attempt of selecting correctly the word of a pair synthesized by the system. Binary value (true, false) indicating whether the user chooses the word of the minimal pair that the system synthesizes in the activity.
- **Discrimination success rate**: percentage of right discrimination attempts according to the total number of attempts of the user in the competition.
- **Number of matches** (Playing or Training mode) in which the user participates in (either launched or answered matches).
- **Match duration**: time a user spends on performing the activities of a match.
- **Challenge win rate**: number of challenges won by a player divided by the total number of challenges in which the user participated in.
- **Leaderboard position, rank**: place on the competition's leaderboard that a player occupies during a challenge.
- **Number of points** obtained from a finished challenge in the Playing mode. The final amount of points achieved by a player in a challenge depends on the

Condition			*ExtraScore*	
IsCreator	MaxBaseScore	BetterRank	$n = 2$	$n \in \{3,4,5\}$
True	True	True	0	0
True	True	False	$rank1 - rank2$	$3 * (n-1)$
True	False	True	$-(rank1 - rank2)$	$-(n-1)$
True	False	False	0	0
False	True	True	0	0
False	True	False	$rank1 - rank2$	$(n-1)$
False	False	True	0	0
False	False	False	0	0

TABLE 7.29: Extra points scoring system of COP (*ExtraScore* value). Table adapted from [31].

performance in her/his corresponding match and the rest of the players. It can be defined as:

$$MatchScore = BaseScore + ExtraScore \tag{7.10}$$

The *BaseScore* is computed from the performance results of discrimination and pronunciation activities in each player's match:

$$BaseScore = \sum_{D=1}^{6} u_D + \sum_{P=1}^{6} v_P; \qquad u_D, v_P \in \{\alpha, \beta, \gamma\} \tag{7.11}$$

where u_D and v_P are the weight values assigned to the activity performance value according to the result: α is the value assigned to a wrong attempt (0), β is the value referred to a right attempt with some help, such as a request for a word listening or performing more than one production attempt (1), and γ is the value assigned to a right attempt without help (2).

The *ExtraScore* is added after all players in the challenge finish their matches. As shown in Table 7.29, this value depends on the number of players, the player who launches the challenge (*IsCreator*) and the leaderboard position difference between the player and the opponents (*BetterRank*) and the *BaseScore*. *IsCreator* value is *True* when the player launched the challenge, *MaxBaseScore* value is *True* when the *BaseScore* achieved by the player is the highest one of the challenge, and *BetterRank* indicates if the leaderboard position of the player is higher than the position of the opponent(s). $rank1$ and $rank2$ are the leaderboard position of the player with the higher and lower position of the leaderboard, respectively. n is the number of players in the challenge.

In particular, the extra points scoring system had the premise of rewarding courageous players. Those who challenged players above in the leaderboard and won, obtained more extra points (no penalties in case of losing the challenge). However, top players were penalized when they challenged worse ones in terms of leaderboard position and lost the challenge.

Proficiency Improvement

The learning improvement analyzed was related to the perception and production skills involved in the activities of the competition. Inter and intra-group success

rates were compared of the same quantity of activities at the beginning (two first days) and at the end (two last days) of the competition.

User Grouping

The total number of Playing matches performed by each user is used to classify COP participants by three statistical tertiles in terms of their quantitative level of activity (performance): T1 (Constant), T2 (Habitual), and T3 (Casual), that is, high, medium, and low participation in the competition, respectively.

7.5.7 Results

Participants of the TipTopTalk! prototype gradually lost interest, most probably due to habituation and lack of new motivational factors. This led us to analyze the effects of a challenge-based competition on user's motivation, performance, and learning (RQ3). In comparison to our previous challenge-free version of the game, TipTopTalk!, the COP challenge-based competition ensured a higher and more stable level of motivation, while also providing a measurable increase in correct pronunciation of the phonemes addressed in the game. Both prototypes shared main gamification elements, such as leaderboards, points, profile avatar, badges, and performance graphs. The results obtained in this experiment, along those derived from the three previous ones, also reinforced the answers to the research questions about the inclusion of current speech technology in a CAPT system (RQ1) and the user's pronunciation performance and learning with a specific pronunciation training methodology (RQ2); by following the research objectives RO1, RO2, RO3, and RO4.

The most relevant results can be classified into four different categories. First, results related to user's behavior with the system (i.e., days of active participation and preferred type of activity). Second, learner's performance interacting with the CAPT system during the competition. Third, the results of the post-quest questionnaires. Finally, results from the focus group sessions. Some of these results have been published in [31]. Their discussion is included in Chapter 8.

User's Behavior

Concerning the distribution of any player's activity throughout the 24 competition days, Figure 7.23 shows the accumulative number of days in which some players' activity was traced (as it was shown in Figure 7.8 for the TipTopTalk! prototype of the second experiment). There was a 17% of the total 354 subjects of the experiment who only participated one day; whereas a 25% played more than 11 days and only a 5% of learners were active during the whole 24 day competition period.

In Figure 7.24 the total distribution of challenges in which each participant on average was involved during the competition is represented in grey bars, whereas the distribution of average number of challenges performed by a user per day is displayed in the horizontal black line. Although the global activity intensity registered falls along the days of competition as the number of participants, the average of challenges per active player remains constant ($\approx 4.5\%$).

User's Performance

Most of user's data gathered was related to discrimination and production events on the Playing and Training modes. Table 7.30 represents the average number of

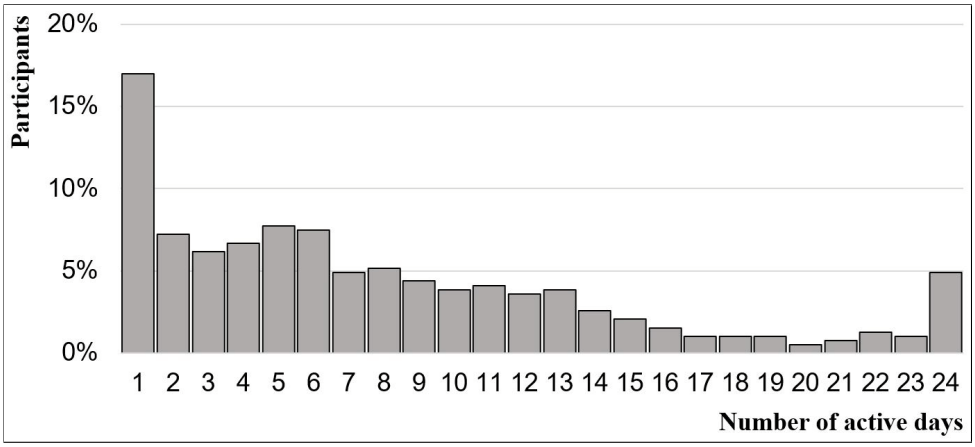

FIGURE 7.23: Distribution of users by number of days with active participation in the COP competition, adapted from [31].

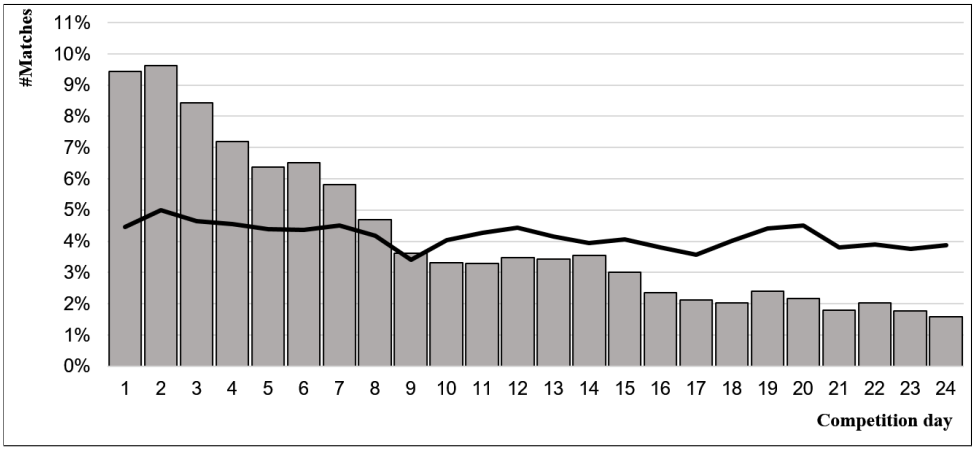

FIGURE 7.24: Distribution of activity registered each day of the COP competition.

these two activity events performed by each user in the experiment. In the Playing mode, an average user performed about six times more production activities than an average player of the TTT prototype (2168.0 vs. 349.9, see Table 7.10) and three times more discrimination activities (1413.2 vs. 405.2). A Mann–Whitney U test shows statistically significant differences in both cases ($U = 186.0$, $p < 0.001$ for productions and $U = 907.0$, $p < 0.001$ for discriminations). In the Training mode, an average user of COP prototype performed almost four times more production activities (81.3 vs. 24.3) and two times more discrimination activities (72.1 vs. 37.0) than an average TTT user. However, a Mann–Whitney U test indicates there were statistically significant differences only in production training activities ($U = 606.5$, $p = 0.022$).

Table 7.31 shows the average values of user's performance indicators interacting with the CAPT system (described in Section 7.5.6). Individuals were categorized by their English level declared in the pre-quest (native or non-native: A1–A2, B1–B2, C1–C2). Since data in Table 7.31 did not pass the Kolmogorov–Smirnoff nor Levene's standard tests, several non-parametric tests for non-normally distributed data were carried out, in order to detect statistically significant differences. In particular, in Table 7.32 the results of a Kruskal–Wallis test [200] conducted to determine the possible statistically significant differences among the three non-native groups (C1–C2, B1–B2, and A1–A2); whereas in Table 7.33, statistical pairwise comparisons were

	$\overline{\#Events}$	$\#Participants \vert \#Total$
Training mode		
Discrimination	72.1 (47.0%)	108/128 (84%)
Production	81.3 (53.0%)	102/128 (80%)
Playing mode		
Discrimination	1413.2 (39.5%)	165/165 (100%)
Production	2168.0 (60.5%)	165/165 (100%)

TABLE 7.30: Average number of discrimination and production events per participant of the COP prototype. The third column ($\#Participants \vert \#Total$) refers to the number of subjects who perform these activities (first value) and the total number of participants who perform an activity of the same mode (second value). Table adapted from [31].

	C1–C2 (48)		B1–B2 (250)		A1–A2 (49)		Non-native (347)		Native (2)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Playing mode										
Production success rate	66.8%	11.6	55.6%	13.4	45.0%	11.6	55.7%	14.1	89.4%	9.4
Discrimination success rate	79.8%	7.0	74.2%	9.7	69.8%	7.9	74.3%	9.5	95.3%	9.4
Number of matches	205.6	229.4	130.4	168.1	80.4	82.3	133	172.0	198	9.4
Match mean duration (s)	73.0	17.3	90.2	39.4	102.4	26.0	89.5	36.3	43.8	9.4
Leaderboard-related										
Challenge win rate	53.0%	22.2	44.4%	30.0	22.4%	16.4	43.6%	23.9	88.1%	9.4
Mean position	141.5	108.6	175.1	101.8	213.0	80.7	175.79	101.483	157.0	9.4
Mean number of points	13.5	5.5	11.5	5.1	9.5	6.3	11.4	5.4	21.51	9.4
Training mode										
Production success rate	29.1%	36.0	32.9%	35.9	32.9%	31.3	32.4%	35.7	-	-
Discrimination success rate	42.1%	41.8	41.5%	36.5	39.5%	38.6	41.3%	37.4	-	-
Number of matches	13.5	22.9	19.0	11.9	8.8	16.0	11.7	21.1	-	-
Match mean duration (s)	37.7	34.4	38.1	31.1	49.6	34.6	40.0	32.2	-	-

TABLE 7.31: Indicators of activity per declared level of English (CEFR) of the COP prototype. SD is the standard deviation. The symbol - stands for not applicable.

carried out determined by Mann–Whitney U tests.

Table 7.32 reports statistical differences for all indicators ($p < 0.05$) except for the Training mode's production success rate ($H(2) = 2.592$, $p = 0.274$), discrimination success rate ($H(2) = 0.677$, $p = 0.713$), and number of matches ($H(2) = 28.507$, $p = 0.395$). In particular, the A1–A2 students were the least implicated ones in the competition with an average of 80.4 matches (Table 7.31). This quantity value was almost three and two times less than the activity registered by the C1–C2 and B1–B2 players (205.6 and 130.4 matches on average).

The C1–C2 players were the most efficient ones since they achieved the highest rates of wins. These values were statistically significant different comparing group by group (see Table 7.33). Regarding the success rate in production and discrimination activities of the Playing mode, the C1–C2 players were also, on average, the most skilled ones, being the differences statistically significant higher than the other groups (see Table 7.33). These differences in the performance of each player have an

		H	df	p
Playing mode	Production success rate	29.997	2	< 0.001*
	Discrimination success rate	27.826	2	< 0.001*
	Number of matches	12.329	2	0.002*
	Match mean duration (s)	21.400	2	< 0.001*
Leaderboard-related	Challenge win rate	23.530	2	< 0.001*
	Mean position	21.754	2	< 0.001*
	Mean number of points	24.439	2	< 0.001*
Training mode	Production success rate	2.592	2	0.274
	Discrimination success rate	0.677	2	0.713
	Number of matches	28.507	2	0.395
	Match mean duration (s)	7.183	2	0.028*

TABLE 7.32: Kruskal–Wallis test results of indicators of activity per declared level of English of Table 7.31 of the COP prototype. The * symbol means that there were statistically significant differences.

		C1–C2 vs. B1–B2	C1–C2 vs. A1–A2	B1–B2 vs. A1–A2
Playing mode	Production success rate	(1062.5, p = 0.005)	(59.5, p < 0.001)	(474.5, p < 0.001)
	Discrimination success rate	(971.5, p = 0.001)	(56.0, p < 0.001)	(596.5, p < 0.001)
	Number of matches	–	(140.5, p = 0.002)	(746.0, p = 0.005)
	Match mean duration (s)	(1200.5, p = 0.033)	(89.5, p < 0.001)	(573.5, p < 0.001)
Leaderboard-related	Challenge win rate	(1204.0, p = 0.034)	(86.0, p < 0.001)	(521.5, p < 0.001)
	Mean position	(1219.0, p = 0.041)	(98.0, p < 0.001)	(543.0, p < 0.001)
	Mean number of points	(1164.5, p < 0.020)	(76.5, p < 0.001)	(531.0, p < 0.001)
Training mode	Production success rate	–	–	–
	Discrimination success rate	–	–	–
	Number of matches	–	–	–
	Match mean duration (s)	–	(182.5, p = 0.024)	(853.0, p = 0.029)

TABLE 7.33: Mann–Whitney U test results (U, p) by declared level of English of Table 7.31 in the COP prototype.

impact on the final positions of the leaderboard: C1–C2 players occupied, on average, higher leaderboard positions than B1–B2, and B1–B2 positions were, at the same time, higher than A1–A2 ones: 141.5, 175.1, and 213.0, mean positions, respectively (see Table 7.31). These differences were statistically significant in the three cases (see Table 7.33). However, only eight C1–C2 players reached one of the first 25 positions on the leaderboard, being the 17 positions remaining occupied by B1–B2 learners. These leaderboard positions aligned with the mean number of points obtained per match: 13.5 vs. 11.5 vs. 9.5, respectively (see Table 7.31). There were significant differences in all groups (see Table 7.33). Finally, there was also a last indicator which also evidences a higher proficiency of C1–C2 players than the rest: the average time spent on Playing matches (73.0s, 90.2s, and 102.4s, respectively, Table 7.31). C1–C2 players spent less time than the other group with statistically significant differences in all cases (see Table 7.33).

Regarding training activities, although B1–B2 players trained, on average, more than the other groups, there were not statistically significant differences in any Training indicator, except for the Training match duration 37.7s, 38.1s, and 49.6s, C1–C2, B1–B2, and A1–A2, respectively, being statistically significant differences between the A1–A2 group and the other two groups (see Table 7.33).

In order to report consistent results, the following results analyzed in this section only include data related to the native Spanish speakers who fully participated in

the experiment (165, see more details in Section 7.5.3). Therefore, we did not include data related to natives, staff members, and native speakers who abandoned at the first stages. These 165 participants fulfilled all questionnaires of the pre/post-quests and completed at least the minimum of 60 challenges for the academic certification. Table 7.34 shows the most relevant activity indicators of these players interacting with the CAPT system, categorized by their involvement on the experiment (number of matches in the Playing mode: Constant, Habitual, and Casual; equivalent to high, medium, and low participation, respectively).

	Constant (56)		Habitual (55)		Casual (54)		Total (165)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Playing mode								
Production success rate	67.4%	9.4	55.5%	11.6	53.2%	9.9	58.8%	12.0
Discrimination success rate	83.3%	4.8	76.0%	5.9	72.2%	6.1	77.2%	7.3
Number of matches	468.7	188.8	144.3	29.7	91.6	11.3	237.1	201.1
Match mean duration (s)	68.3	14.1	85.6	27.6	97.0	65.6	83.5	43.0
Leaderboard-related								
Challenge win rate	49.4%	11.5	43.4%	17.4	41.0%	16.4	44.6%	15.6
Mean position	29.4	17.8	98.3	37	149.2	33.0	91.6	57.8
Mean number of points	17.5	3.0	12.7	3.2	12.2	3.1	14.1	3.9
Training mode								
Production success rate	50.7%	33.4	47.5%	33.4	21.8%	30.4	40.2%	34.8
Discrimination success rate	58.1%	33.6	52.6%	32.4	30.3%	36.3	47.2%	36.0
Number of matches	30.2	35.3	13.6	18.8	6.0	13.5	16.7	26.4
Match mean duration (s)	36.8	17.8	50.9	41.4	39.3	33.9	42.3	32.9

TABLE 7.34: Indicators of activity per type of user of the COP prototype. SD is the standard deviation. Table adapted from [31].

The values of the eleven indicators of the experiment, presented in Table 7.34, were statistically analyzed in Table 7.35 and Table 7.36. In particular, in Table 7.35 the results of a Kruskal–Wallis test conducted to determine whether there were statistically significant differences among the three groups about the indicators are displayed; whereas in Table 7.36 a pairwise comparison determined by Mann–Whitney U tests is presented. Table 7.35 reports differences for all indicators ($p < 0.05$) except for the Training mode's match mean duration time ($H(2) = 3.099$, $p = 0.212$).

In particular, the Casual group was the least implicated one in the competition with an average of 91.6 matches (Table 7.34). On the other hand, the Constant players were very active, reaching 468.7 matches on average. The 20 most active users of this last group performed a mean of 29 matches per day (30 was the maximum allowed). In this group, there were users that competed until the end of the competition to climb up to the top of the leaderboard. Although having a few opportunities of achieving the rewards, the Habitual group was composed of users who keep on playing after reaching the minimum to obtain the academic certification. Table 7.36 shows statistically significant differences in the three cases.

The Constant players were the most efficient ones since they achieved the highest rates of wins and production success. These values were statistically significant different with respect to the other two groups (see Table 7.36). Regarding the success rate in discrimination activities, they were also the most skilled ones, the differences being statistically significant higher than the other groups (see Table 7.36). These

		H	df	p
Playing mode	Production success rate	45.413	2	< 0.001*
	Discrimination success rate	73.757	2	< 0.001*
	Number of matches	145.785	2	< 0.001*
	Match mean duration (s)	40.930	2	< 0.001*
Leaderboard-related	Challenge win rate	9.590	2	0.008*
	Mean position	126.213	2	< 0.001*
	Mean number of points	63.029	2	< 0.001*
Training mode	Production success rate	21.289	2	< 0.001*
	Discrimination success rate	18.275	2	< 0.001*
	Number of matches	28.507	2	< 0.001*
	Match mean duration (s)	3.099	2	0.212

TABLE 7.35: Kruskal–Wallis test results of indicators of activity of Table 7.34 in the COP prototype. The * symbol means that there were statistically significant differences. Table adapted from [31].

		Constant-Habitual	Constant-Casual	Habitual-Casual
Playing mode	Production success rate	(659.5, $p < 0.001$)	(449.5, $p = 0.004$)	–
	Discrimination success rate	(512.5, $p < 0.001$)	(168.0, $p < 0.001$)	(987.5, $p = 0.003$)
	Number of matches	(687.0, $p < 0.001$)	(435.0, $p < 0.001$)	(876.0, $p < 0.001$)
	Match mean duration (s)	(746.0, $p < 0.001$)	(514.0, $p < 0.001$)	–
Leaderboard-related	Challenge win rate	(1136.5, $p = 0.017$)	(1028.0, $p = 0.004$)	–
	Mean position	(37.0, $p < 0.001$)	(0.0, $p < 0.001$)	(429.0, $p < 0.001$)
	Mean number of points	(413.5, $p < 0.001$)	(345.5, $p < 0.001$)	–
Training mode	Production success rate	–	(834.0, $p < 0.001$)	(885.5, $p < 0.001$)
	Discrimination success rate	–	(875.5, $p < 0.001$)	(976.0, $p = 0.001$)
	Number of matches	(1196.5, $p = 0.042$)	(702.0, $p < 0.001$)	(839.5, $p < 0.001$)
	Match mean duration (s)	–	–	–

TABLE 7.36: Mann–Whitney U test results (U, p) for the three group pairs of Table 7.34 in the COP prototype. Table adapted from [31].

differences in the performance of the player groups have an impact on the final positions of the leaderboard. Constant players occupied higher leaderboard positions than Habitual, and Habitual positions were, at the same time, higher than Casual ones: 29.4, 98.3, and 149.2, mean positions, respectively (see Table 7.34). The mentioned differences were statistically significant different in the three cases (see Table 7.36). These leaderboard positions aligned with the mean number of points obtained per match: 17.5 vs. 12.7 vs. 12.2, respectively (see Table 7.34). In this case, there were significant differences in two cases: Constant vs. Habitual and Constant vs. Casual (see Table 7.36). There was a last indicator which also evidences a higher proficiency of Constant players than the rest: the average time spent on Playing matches (68.3s vs. 85.6s, and 97.0s, respectively, Table 7.34). Constant players spent less time than the other group with statistically significant differences (see Table 7.36).

Regarding training activities, Constant players trained more than the rest of the users: 30.2 vs. 13.6 vs. 6 (number of Training matches, Table 7.34). These differences were statistically significant in all cases (see Table 7.36). In the case of production and discrimination events, the Constant group players achieved higher success values than the Habitual group ones, and Habitual players achieved better values than the Casual players (see Table 7.34). The differences were statistically significant in both cases when comparing Constant vs. Casual, and Habitual vs. Casual (see Table 7.36).

Player's perception and production improvement rates at the beginning and at

		Constant (56)		Habitual (55)		Casual (54)		Total (165)	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Production	First (%)	66.7	7.8	53.7	8.8	52.9	7.9	57.8	10.3
	Last (%)	73.4	5.7	56.1	10.4	55.2	8.0	61.7	11.8
	Diff (%)	6.7	6.5	2.4	8.4	2.3	5.5	3.8	7.2
	(Z, p)	(-5.7, < 0.001)		(-3.2, 0.001)		(-2.9, 0.004)		(-7.1, < 0.001)	
Discrimination	First (%)	76.9	5.7	73.7	7.3	71.8	6.2	74.1	6.7
	Last (%)	87.0	5.7	78.1	6.7	73.6	7.9	79.7	8.8
	Diff (%)	10.1	4.7	4.4	6.4	1.8	6.6	5.4	6.9
	(Z, p)	(-6.5, < 0.001)		(-4.3, < 0.001)		(-2.9, 0.003)		(-8.6, < 0.001)	

TABLE 7.37: Success rates of discrimination and production events at the beginning and at the end of the COP prototype. *First* and *Last* represent the right success rates during the first and last two days of the competition. *Diff* is the absolute difference between *First* and *Last*. Z and p are values derived from a Wilcoxon signed-rank test at 95% confidence level (2-tailed). *SD* is the standard deviation. Table adapted from [31].

the end of the competition are displayed in Table 7.37. Particularly, the activities performed during the first and last two days of the competitions were measured. In particular, it compares the users' performance at the beginning and end of the experiment, classified by tertiles. Improvements were observed in the three groups (positive differences between the columns *First* and *Last*, both in production and discrimination). A Wilcoxon signed-rank test shows that these differences were statistically significant in all cases in production activity types ($p < 0.001$, $p = 0.001$ and $p = 0.004$ for Constant, Habitual and Casual, respectively), and in discrimination ones ($p < 0.001$, $p < 0.001$ and $p = 0.003$ for Constant, Habitual and Casual, respectively).

The most active players in the experiment (users in the Constant group) improved the most in production: 6.7% vs. 2.4% (Habitual) and 2.3% (Casual). They also improved the most in discrimination: 10.1% (Constant) vs. 4.4% (Habitual) and 1.8% (Casual). There was a statistically significant difference among player types, both in production activities ($H(2) = 11.745$, $p = 0.003$, Kruskal–Wallis test) and in discrimination ($H(2) = 47.507$, $p < 0.001$, Kruskal–Wallis test). A Mann–Whitney U test was conducted to determine whether there was a statistically significant difference between the user groups. In production, this test indicates that there was a statistically significant difference between Constant and Habitual ($U = 1088.5$, $p = 0.008$); and between Constant and Casual ($U = 982.5$, $p = 0.002$). In discrimination, a Mann–Whitney U test indicates that there were statistically significant differences in two cases: between Constant and Habitual ($U = 737.0$, $p < 0.001$); and between Constant and Casual ($U = 380.0$, $p < 0.001$).

Post-quest Questionnaire

Tables 7.38 and 7.39 shed light on the reasons behind the different users' level of activity and behavior during the competition. The majority of players declared that improving their English pronunciation (Q1.3 75.2%) and climbing up the leaderboard (Q1.6 70.3%) were the principal reasons for playing. Winning a prize (Q1.8 48.5%) and beating a known playmate (Q1.7 42.4%) were the third and fourth reported reasons.

Question	Casual	Habitual	Constant	Total	Answer (165)
Q1.1	1.9%	0.0%	0.0%	0.6%	- (No answer)
Q1.2	11.1%	16.4%	37.5%	21.8%*	I was hooked on the game, and played as much as I could.
Q1.3	74.1%	80.0%	71.4%	75.2%	The prospect of improving my English pronunciation was an important incentive in using the app.
Q1.4	20.4%	18.2%	5.4%	14.5%*	I felt obliged to play, basically because I needed the academic certificate.
Q1.5	18.5%	21.8%	26.8%	22.4%	I have played a lot because I felt that my pronunciation was getting better.
Q1.6	50.0%	69.1%	91.1%	70.3%*	Climbing up the leaderboard was my main incentive in using the app.
Q1.7	50.0%	41.8%	35.7%	42.4%	The possibility of challenging my mates was an important incentive for using the app.
Q1.8	20.4%	47.3%	76.8%	48.5%*	Winning the prize was an important incentive for using the app.

TABLE 7.38: Declared reasons for playing of the COP prototype. The question in the questionnaire is: *Select the statements that fit your motivation for playing; you can choose as many as you consider appropriate*. The symbol * indicates statistically significant inter-group differences Chi–square test at 95% confidence. Table adapted from [31].

Concerning the inter-group differences, winning a prize (Q1.8) was a predominant engaging reason for playing for Constant players: 76.8%, being clearly higher than in the other two groups Habitual and Casual, 47.3% and 20.4%, respectively. There were statistically significant differences in all cases: between Casual and Habitual ($\chi^2(1) = 8.795$, $p = 0.003$); between Casual and Constant ($\chi^2(1) = 35.010$, $p < 0.001$); and between Habitual and Constant ($\chi^2(1) = 10.275$, $p = 0.001$). Few players (less than 21% in any case) declare the academic certificate was the principal reason (Q1.4); this have less impact for Constant players: 5.4%, with a statistically significant difference with respect to Casual players ($\chi^2(1) = 5.579$, $p = 0.023$).

The motivation for climbing up the leaderboard (Q1.6) made the difference: 91.1% of Constant players vs. 69.1% (Habitual) and 50.0% (Casual). There was a statistically significant difference between Constant and Casual ($\chi^2(1) = 22.481$, $p < 0.01$) and between Constant and Habitual ($\chi^2(1) = 8.436$, $p = 0.04$). The percentage of players that declare themselves "motivated players" (Q1.2) reaches 37.5% in the Constant group, this percentage being lower for the rest of the groups: 16.4% and 11.1%. There was a statistically significant difference between Constant and Casual ($\chi^2(1) = 10.337$, $p < 0.002$) and between Constant and Habitual ($\chi^2(1) = 6.285$, $p = 0.018$).

None of the possible answers depicted in Table 7.39 were over 40% (except for the Habitual answer for Q2.7), and in fact, close to 40% of the users do not declare any negative opinion toward competition (Q2.1). Only a 3% of the players declare more enjoyment during training than in competition (Q2.2) and again only 3% of players declare having suffered anxiety during the game (Q2.6). Concerning inter-group differences, Constant users report feeling uncomfortable with a lot quantity of matches (Q2.5) 28.6% vs. 10.9% and 13.0%. There was a statistically significant difference between Constant and Casual ($\chi^2(1) = 4.050$, $p = 0.044$) and between Constant and Habitual ($\chi^2(1) = 5.447$, $p = 0.020$). The answer related to going at their own will (Q2.7) are also interesting, not only because it was the most selected answer overall, 39.4%, but also because it made the difference between Constant (23.2%) and the other groups Habitual (51.9%) and Casual (43.6%). There was a statistically significant difference between Constant and Casual ($\chi^2(1) = 9.643$, $p < 0.002$) and between

Question	Casual	Habitual	Constant	Total	Answer (165)
Q2.1	31.5%	40.0%	44.6%	38.8%	- (No answer)
Q2.2	3.7%	5.5%	0.0%	3.0%	I enjoyed the training matches more than the pressure of the game.
Q2.3	16.7%	12.7%	10.7%	13.3%	I would rather play against the machine and against myself privately.
Q2.4	7.4%	5.5%	1.8%	4.8%	Appearing in a leaderboard and challenging others makes me feel bad and puts me off participating.
Q2.5	13.0%	10.9%	28.6%	17.6%*	I feel uncomfortable with so many challenges.
Q2.6	1.9%	3.6%	3.6%	3.0%	The challenges have caused me anxiety and discomfort.
Q2.7	51.9%	43.6%	23.2%	39.4%*	I like to advance at my own pace without comparing myself with others.
Q2.8	7.4%	5.5%	5.4%	6.1%	It makes me uncomfortable to find myself in a situation where I have to prove that my pronunciation is good.

TABLE 7.39: Attitude toward competition of the COP prototype. The question in the questionnaire is: *Check the statements expressing your feelings toward competition during the game; you can choose as many as you consider appropriate*. The symbol * indicates statistically significant inter-group differences Chi–square test at 95% confidence. Table adapted from [31].

Habitual and Constant ($\chi^2(1) = 5.208$, $p = 0.022$).

Question	Percentage	Answer (129)
Q3.1	42.4%	Technical reasons (*my mobile device is not Android, I cannot install the game...*).
Q3.2	41.7%	Lack of time.
Q3.3	27.3%	It bothered me that other users did not accept the challenges or the delayed time until accepting them.
Q3.4	19.0%	The game bothered me, I did not like it.
Q3.5	12.9%	I did not find it useful to learn English.
Q3.6	4.6%	I was frustrated not being able to win.
Q3.7	3.8%	The competition made me feel bad.

TABLE 7.40: Early abandonment questionnaire results of the COP prototype. Table adapted from [31].

Finally, 129 out 182 players who quit the experiment before completing all mandatory protocol steps answered a questionnaire about reasons for early abandonment. Table 7.40 shows that most early abandonment reasons were technical reasons (Q3.1 = 42.4%) and lack of time (Q3.2 = 41.7%). Furthermore, few players reported to be uncomfortable during the competition (Q3.7 = 3.8%) or frustrated because of losing (Q3.6 = 4.6%).

Focus Group Sessions

Four focus group sessions were carried out with 16 different non-native learners of the experiment in each one of them (64 in total) who were randomly selected according to four characteristics obtained from the specific questionnaires of the prequest: extrinsic and intrinsic motivation, degree of competitiveness, and different English proficiency levels.

The notes extracted from the subjects participation of the two first focus group sessions are summarized in Table 7.41. They involved answers from participants

	Focus group session (1 and 2)	
	Intrinsic motivation	**Extrinsic motivation**
Strategy	*"There were people who did not accept my challenges and I could not go forward in the competition". "I mainly submitted challenges". "First, I accepted all incoming challenges. Then I submitted the remaining ones". "I ended the competition challenging the same group of participants". "The first challenges of the competition were a disaster because so many challenges were ignored". "When I saw myself in low positions on the leaderboard, I quit playing".*	*"I tried to answer all incoming challenges". "I only challenged users above my position on the leaderboard". "When a user accepted my challenges I submitted another one". "My strategy consisted in finding users who accepted my challenges instead of trying to improve my pronunciation". "Climbing up the leaderboard was very motivating". "I played because I enjoyed the game".*
Training	*"I think people did not train because it does not reward points". "I found the Training mode very interesting". "I trained words that I did not utter correctly".*	*"I did not train because I did not get points". "I preferred playing than training". "I felt bored training".*
Level	*"I thought I was better at English". "I realized the differences between similar words". "I think I am now capable of distinguishing more English vowels". "When you listen to the minimal pair words consecutively, you realize their differences". "The phonetic transcription helped me so much". "I think this app is very interesting to play at class".*	*"I think I have improved my pronunciation". "At the end of the day I could produce words better". "This game allows me to improve my pronunciation".*
Others	*"Similar apps only focus on grammar or vocabulary. However, this app could actually recognize my voice". "At the beginning I was ashamed to produce words near to my roommates. Some days after I did not mind". "I felt frustrated sometimes because the CAPT tool did not accept several words".*	*"I tried to play in silent places". "The voice recognizer was sometimes unsettling".*

TABLE 7.41: Notes gathered from the intrinsic and extrinsic focus group sessions of the COP prototype.

with a high declared intrinsic motivation level (first column) and a high extrinsic declared one (second column), categorized by the topics presented in the session (strategy followed during the competition, training activities value, implication on the English proficiency level, and others). In particular, intrinsic motivation refers to do something because it is personally rewarding to yourself; whereas extrinsic motivation involves doing something because you want to earn a reward. Questions in these two sessions were mainly focused into the strategy developed during the competition and the reasons for keep on participating. Answers related to the strategy followed for challenging other users were clearly differentiated. Extrinsically motivated users preferred to be sure that their challenges are completed to obtain the maximum possible number of points with an evident strategy of climbing up on the leaderboard. Intrinsically motivated players pointed out the value of training activities and the possibilities that the CAPT system offers to improve their pronunciation.

These results were aligned to those presented in Table 7.30 in which the number of training activities was significantly lower, from a statistical point of view, than the number of playing ones; with results from Table 7.34 in which the quantity of training matches was significantly lower, from a statistical point of view, than the number of playing ones; and with the answers given to questionnaire about reasons

for playing from Table 7.38.

	Focus group session (3)
	English level
Strategy	*"If your challenges were ignored you could not go forward". "In general, I thought it was a competition and not learning". "I tried to challenge users above myself in the leaderboard". "I ignored most challenges from users who were below myself in the leaderboard".*
Training	*"I only trained when there were words I did not understand well". "Training allowed myself to realize how the app produced the words". "I needed to train for climbing up to the first position on the leaderboard" "I preferred to play the 30 challenges instead of training".*
Level	*"I think this app helps you to pronounce better". "I appreciate pronunciation-oriented educational apps". "I thought I was not able to produce so many different sounds, but this app helped me". "It is very useful to become aware of some sounds". "This type of application is very useful to realize that in English there are more than five vowels". "I think it is very positive you can challenge to others in the competition". "I find it very positive to learn the pronunciation of English in general". "It is basically fun and helps you to be aware of what you have not heard before". "If you train with a machine that recognizes English, you are also training how you have to speak as an English native". "I would recommend the application for people who are intermediate in English". "The phonetic transcription was useful to see how the word was pronounced". "I did not realize the sound was synthetic". "If there was noise in the environment it did not work well".*
Others	*"I would add the definition or the translation of the words". "It was exasperating that the recognizer did not recognize the word despite saying it well". "I would love to try it in British English".*

TABLE 7.42: Notes gathered from the English proficiency level focus group session of the COP prototype.

Subjects who reported a high L2 English proficiency level in the pre-quest participated into the third focus session. The main topic discussed was related to the possible impact of the CAPT system into the pronunciation level improvement (see Table 7.42). Their positive opinions about the influence of the CAPT system in user's English proficiency level were aligned with results of Table 7.37, in which an improvement in perception and production skills were reported. These results also shared similarities to those presented in Table 7.39, about the user's attitude toward competition.

Finally, the last focus group session involved subjects who declared a different degree of competitiveness in the pre-quest (see Table 7.43). User's opinions in this group varied from players who played to obtain a final reward, to those who played just for fun or to feel motivated. In accordance with other focus group sessions, training activities were reported to be used very scarcely. These results agreed with those presented in Table 7.34, in which the quantity of training matches was lower than the number of playing ones, from an statistically significant point of view. They also shared similarities to those presented in Table 7.38, about the reasons for keeping on playing.

7.6 Summary

In this chapter, the four main experiments carried out in this thesis have been reported. All of them involved participation of real users and the informed design and development of mobile native applications needed to conduct them. The theoretical

	Focus group session (4)
	Degree of competitiveness
Strategy	*"If someone ignored my challenges I did not accept her/his challenges". "I preferred to challenge those who were above on the leaderboard because they were opened to answer more likely". "There were more possibilities to be ignored in challenges with more than two players". "The points system seemed too complicated". "I stopped playing because I saw that it was impossible to win". "I consider adequate the limit of 30 daily challenges". "I felt I submitted so many challenges and so many were unanswered".*
Training	*"I trained because I could listen to the words I found impossible to pronounce correctly". "I liked to be able to choose specific phonemes to practice in the Training mode" "I trained a little, only when I had problems with some words"*
Level	*"I would play, even if there was no reward, just to learn". "I felt motivated when I won". "I think I improved my pronunciation". "I tried to play in isolated places, not only because of the noise, but also the shame". "My motivation was to achieve the academic certification". "I have learned some phonetics". "I believe producing isolated words can help me to produce sentences". "I liked to reach more points at the end of the competition than at the beginning". "If there had been no incentive, I would not have played or I would have played less".*
Others	*"I think it would be a great idea to contextualize the words". "I felt overwhelmed when I needed to pronounce so many times". "I would like to skip words when I have problems".*

TABLE 7.43: Notes gathered from the degree of competitiveness focus group session of the COP prototype.

concepts, practical approaches, and training strategies introduced in the state-of-the-art and the experimental framework have been adapted to the experimentation. Volunteer speakers with different levels of pronunciation skills and coming from several academic institutions have participated in the experiments, allowing us to gather a significant amount of spoken, interaction, and subjective evaluation data. To avoid an excess of dispersion in the mainstream of the document, a comparative view of the main characteristics of the experiment is described in Appendix B.

The first experiment, *Alpha*, provided a starting point to check ASR, TTS, and minimal pairs suitability for the first CAPT tool prototype developed, *Minimal Pairs*. There were several limitations since native speakers did not successfully complete all the pronunciation activities. The position of the target word in the ASR hypotheses list and the g-score matched with the declared level of pronunciation proficiency of the speakers. Non-native speakers reported useful the isolated use of the TTS when failing production attempts. Finally, the opinions in the focus group session about the CAPT tool, suggesting new activities, feedback techniques, and motivational elements were taken into account for the next experiments.

In the second experiment, *Non-guided Learning*, we targeted measuring the improvement of production and discrimination abilities by users along a one month period during which an individualistic approach to social competition was adopted, using *TipTopTalk!*, a gamified CAPT system. Discrimination activities were the most performed ones, leading to a general improvement in discrimination skills. However, despite the introduction of gamification elements and the mentioned improvement in discrimination, a stagnation in training intensity and production improvement was detected, being higher in the best players.

These problems led to a third experiment, *Guided Learning*, which moved to a guided training protocol and removed the gamification elements. The CAPT tools developed, *English Vowels* and *Japañol* included recommended activities based not

only on users' L1 and L2, but also on their results, and offered a more specific and personalized feedback than the previous experiments. Results showed a significant pronunciation improvement among the learners who trained with the CAPT tool, and a correlation between human rater's assessment of post-tests and automatic CAPT assessment of users.

Finally, and almost in parallel to the previous one, the fourth experiment, *Competitive Learning*, analyzed the implications on user's motivation, performance, and learning outcomes of a challenge-based competition using a CAPT tool, *COP*. This tool is the second version of the *TipTopTalk!* CAPT tool, which shared the same gamification instruments and activities, but including a competitive scenario in which players had to "challenge" other participants via pronunciation activities. Results showed intensive practice supported by a significant quantity of activities and playing regularity, so the most active and motivated players in the competition achieved significant pronunciation improvement results.

Chapter 8

Discussion

In this chapter, we retake the research questions defined in Section 1.3 to guide the critical revision and evaluation of the experimental results presented in previous chapters, showing their importance and how they relate to the results found in the literature.

The discussion is divided into three main parts, according to the research questions of this thesis, defined in Section 1.3. Each part includes not only the discussion of the results presented in the previous Chapter 7, but also their limitations. First, the feasibility of integrating current TTS and ASR technologies into CAPT systems is examined. Second, the different training methodologies for pronunciation training with CAPT systems applied to the experiments, focusing on the feedback and assessment provided, are discussed. Finally, the effect of game elements and approaches on user's performance, motivation, and learning outcomes are analyzed.

8.1 TTS and ASR Technology in CAPT Systems

The first research question, **RQ1**, addressed the use of current TTS and ASR technologies as non-obstructive pedagogical resources within the CAPT systems developed; whereas the ***Issue 1.1***, was concerned with finding out whether the TTS and ASR technologies integrated into the CAPT systems developed could help to assess different L2 pronunciation level of learners.

The inclusion of TTS technology did not obstruct learners to perform pronunciation activities in the experimentation (RQ1). Both, the TTS and ASR technologies integrated into the CAPT systems developed in this thesis were inserted in a well-defined pedagogical background. Non-native learners resorted to the TTS system when faced to misproductions. In the Alpha experiment the number of interactions reached statistically significant differences inasmuch as the declared proficiency level of the speaker was lower (see Table 7.7). In addition to the production tasks, more training activities such as exposure and discrimination, in which users needed to interact directly with the TTS system to successfully perform them, were included into the competitions of the second and fourth experiments, Non-guided Learning and Competitive Learning, respectively (see Tables 7.11 and 7.34). According to the results reported in both prototypes of the Guided Learning experiment, the TTS system used for generating pronunciation models helped learners to improve their perception and production skills, being also fully functional as feedback in the exposure mode. In particular, students in the experimental group frequently utilized the synthesizer when faced with difficulties both in discrimination and production

activities, as the previous experiments (see Tables 7.14 and 7.15 for the English Vowels prototype, and Tables 7.22 and 7.23 for the Japañol prototype).

TTS technology also helped learners to improve their pronunciation, both, objectively with data gathered with the CAPT systems, and subjectively according to the opinions gathered in the focus group sessions (RQ1). The TTS system employed appeared to be beneficial for students: success rate values significantly increased after undertaking the exposure activities imposed by feedback (vs. not following the feedback), both in the number of requested listenings (see Tables 7.18, 7.24 and 7.25) and in the exposure activities recommendations in the prototypes of the Guided Learning experiment (see Table 7.17). The quality of the sound generated by the TTS to perform the proposed activities not only helped, oriented and assisted players of the COP prototype to successfully perform the activities (see Table 7.34), but also, the TTS was highly valued in the focus group sessions ("*When you listen to the minimal pair words consecutively, you realize their differences*", "*I did not realize the sound was synthetic*", "*I trained because I could listen to the words I found impossible to pronounce correctly*", see Tables 7.41, 7.42, and 7.43).

Using TTS technology in CAPT systems saved human and money costs in the pronunciation training experimentation since no human voice was needed (RQ1). Although it cannot be stated that the quality of the TTS was, by itself, responsible for the pronunciation improvement reported in the experiments, it can be pointed out that the natural process of learning does not have a lack of quality [97] and speech synthesis can be perceived as good as native voice [72]. Since the process of traditional approaches to rely on a specific and proper set of natural speech utterances for any experiment (recorded or live) is not trivial and in much cases, very complicated and expensive to carry out, the feasibility of using TTS technology leads to an innovative resource to future CAPT projects and has clearly evidenced to be non-obstructive to the process of learning. This fact is worth considering since pronunciation improvement results obtained in the experiments were positive and comparable to those obtained by in-classroom training groups, when appropriate.

The inclusion of ASR technology in CAPT systems did not obstruct learners to perform pronunciation activities in the experimentation (RQ1). The purpose of an ASR system is, in fact, equal or even more important than the TTS one since it offers diagnosis as automatic feedback, being non-obstructiveness also a relevant issue. Although it must be taken some precautions in their integration, ASR systems in their present state lend feasibility to CAPT projects. A well-designed CAPT should not be affected by shortcomings, such as environmental noise when recognizing speech, as long as the ASR system had been tested properly with the target words. As it will be discussed in Section 8.1.1, it has been necessary to carefully explore and test the potential of the ASR systems with a specific protocol to add/discard elements which are neither possible to be recognized nor synthesized by the speech technology (see Section 6.1.2).

ASR technology also helped to report learner's pronunciation improvement, both, objectively with data gathered with the CAPT systems, and subjectively with the learner's opinions in the focus group sessions (RQ1). In general, students who interacted with the CAPT systems of all experiments of this thesis performed a high number of profitable interactions with the ASR technology included, reporting improvement in production activities (see Tables 7.19, 7.28, and 7.37). Besides, the positive user's opinions gathered from the focus group sessions of the Alpha and

Competitive Learning experiments about the useful role of the ASR system supported this idea ("*The answer given by the tool in each utterance was very fast*", "*In general, I think this tool could be useful to improve my pronunciation with more sounds*", "*This game allows me to improve my pronunciation*", "*At the end of the day I think I could produce words better*", see Sections 7.2.6 and 7.5.7, respectively). Although the quality of the ASR technology integrated into the CAPT systems was not analyzed as an independent variable, it was enough not to constitute an impediment to pronunciation improvement.

The inclusion of TTS and ASR technology in CAPT systems helped to assess different L2 pronunciation level learners (*Issue 1.1*). In order to offer specific types of activities and provide a personalized feedback according to the student's L2 level, results derived from the Alpha experiment were a starting point to confirm that, in fact, the TTS and ASR systems employed could help to assess different groups of speakers by their L2 pronunciation level (***Issue 1.1***): Tables 7.2 and 7.3 displayed statistically significant better results by native participants (less total and average production attempts and more right and success rate values than the advanced level non-natives students of Group B, and they, in turn, better results that the low-level students of Group C). Also, the total and average time devoted to complete the training session was lower inasmuch as the declared proficiency level of the speaker increased (Tables 7.2 and 7.4). In the case of the TTS system, Table 7.7 showed that native speakers barely used it (only when the ASR system did not recognize their utterances); whereas the Group C made a statistically significant use of the TTS system achieving better recognition results after listening to the synthesized models than the rest of groups. These results agreed with those reported in Figure 7.10 of the Non-guided Learning experiment, in which the three groups of user's pronunciation level kept their differences during the whole experiment (24 days). In the case of the COP prototype, as the rest of experiments, non-native speakers from beginner to expert L2 level declared, achieved statistically significant different results, being them better for the higher declared level. Besides, the average match time duration was higher for lower declared level students (see Table 7.31).

Current TTS and ASR technology offered different quality metrics to assess L2 pronunciation in the CAPT systems developed in this thesis (*Issue 1.1*). In particular, the objective measurements obtained in the first experiment, Alpha, were a first approximation to the representation of user's pronunciation quality, and they were improved in the consecutive experiments. The g–score values (see Table 7.4) and the position of the expected word in the list of string hypotheses of the ASR system (see Table 7.5), the time spent to perform the activities (see Table 7.4), the number of times the TTS system is resorted to (see Table 7.7), the success rates of discrimination and productions tasks in specific time windows, and the average value of the success scores obtained by a user in each training mode and lesson of the CAPT system (G, see Section 7.4.10), have been used as indicators of the quality of the pronunciation. The availability of these objective metrics combined with more sophisticated ones, could lead to suggest corrective feedback and training activities to those speakers who achieve unsatisfactory results in further experiments (see Section 8.2).

In summary, even when current TTS and ASR systems are not designed to be used as specialized pedagogical tools but to provide effective voice human-computer interaction, the results described in this thesis and the ones found in the (scarce) literature confirm the hypothesis that these technologies are ready to use in LL activities without obstructing the natural process of learning. In particular,

in this dissertation these speech technology systems have been integrated into the CAPT tools developed for the prototypes of the experiments with detailed experts' knowledge and within carefully designed and research-based teaching protocols. In fact, the key to their effectiveness seems to lie not only in the sequencing of specific training activities offered to the learners, but also in the recommended corrective feedback strategies designed by experts, in which TTS and ASR technologies are included.

8.1.1 Limitations

Along the last decade, there has been a constant improvement of the quality of TTS [7], [81] and ASR [9], [10] systems, both for general use and for LL applications. Nevertheless, after the experimentation carried out along this thesis, we recognize there are still some important problems to address to get better results in future applications to CAPT. In particular, the feasibility of these technologies largely depended on the tasks they have been originally designed for. We have identified 5 types of factors to be taken into account when working and elaborating minimal pairs lists when using ASR and TTS technologies:

- **False alarms**. Natives' success rate values were not 100% in any experiment in which they were involved (Minimal Pairs, TipTopTalk!, and COP prototypes). This problem was also reported in [37], [57].
- **False positives**. In the case of the Google ASR system, it sometimes succeeded at recognizing words which had been deliberately pronounced with transferred pronunciation. For instance, when a native Spanish speaker feeds a totally transferred version of the word "*tool*" to the ASR system —i.e., an expression that, in fact, matches the Spanish word "*tul*"— it is recognized the English word "*tool*" in the first positions of the *n*–best list of string hypotheses. However, the differences between both words in terms of articulation are by no means trivial: Spanish /t/ is dental and lacks aspiration; Spanish /u/ is closer, less dynamic in terms of tongue stability, and more retracted than English /u:/; and Spanish /l/ lacks the velarization that characterizes English /l/ at syllable coda. This evidences that the probabilistic approach used by Google is blinded to such articulatory features and to its acoustic counterparts. Accomplished recognizability does not guarantee native-like accuracy, and it is not even regularly linked to it: the recognizable and the accurate will be at different distances depending on the number of existing probabilistic alternatives to any given input. This problem of transferred pronunciation has been taken into account for the words selection criteria proposed in this thesis (as explained in Section 6.1.2).
- **Homophones**. One of the most important factors to decide if a utterance was correct or incorrect is the knowledge of other words that are produced exactly the same but are written differently. Also for speech synthesis, these words can be interchanged. It has been necessary to count on expert's abilities and specific dictionaries with words and their phonetic transcriptions to find them in all experiments since there could be more than one homophone per word.
- **Word frequency**. ASR systems are generally intended to recognize the most frequent words of a language since training models have less occurrences of infrequent words. Current TTS systems barely undergo this problem since they

can generate almost any word (real or invented) by joining sounds. However, sometimes the final result of the joining is not intelligible enough.

- **OOV words**. There were some words which were penalized when produced in isolation since they are not usually found in natural language as one-word sentences, as this thesis activities required.

8.2 Training Methodology in CAPT Systems

RQ2 addressed the relation between the methodological elements designed, such as the use of exercises based on minimal pairs within a cycle of training activities, and the pronunciation improvement measured after interacting with the CAPT system of the experimentation. The main interaction events between the users and the CAPT system consisted in tasks of listening, discrimination, and production of short elements (words of minimal pairs). The minimal pairs lists included in the CAPT systems followed a specific selection protocol which evolved from the basic manual selection by an expert of short lists to a semi-automatic protocol which helped experts to find all possible minimal pair combinations of a language and which can be adapted to the speech technology (TTS and ASR) of the CAPT system (explained in Section 6.1.2). This process allowed us to elaborate lists of minimal pairs for the experiments, reducing human and time costs and being possible to be applied to almost any language.

The big number of tasks performed by the students during the experimentation became a relevant factor in explaining the improvement results reported at the end of each experiment (RQ2). A high training intensity was confirmed in the two experimental competitions, TipTopTalk! and COP, a user on average listened to 1442.5 and 2484.3 words, respectively, and produced 374.2 and 2249.3 word-utterances (see Table 7.10 and Table 7.30). In the case of the English Vowels and Japañol prototypes (Guided Learning experiment), in which learners participated in three one-hour sessions with an effective and objectively registered time of at least a 55% of the total sessions time, resorted to the TTS system on average 831.2 and 612.6 times (see Table 7.14), respectively, and made use of the ASR system 615.6 and 291.74 times (see Table 7.22), respectively. This quantity of exercises could be also considered to explain the differences between the post-test results of the experimental and in-classroom groups of both prototypes.

In this work, we have found empirical evidences about the fact that methodology related design issues do have a relevant and noticeable impact both on pronunciation improvement by students using our CAPT tools and on their motivation and level of activity (RQ2). The issues related to RQ2 were specifically addressed from the second experiment. Regarding the assessment of the possible improvement in learner's pronunciation after using the CAPT system in this second experiment (***Issue 2.1***), results displayed higher discrimination and production scores at the end of the competition. Students of the COP prototype also achieved statistically significant higher discrimination and production success rate values at the end of the competition (see Table 7.37). Besides, both prototypes of the third experiment (Guided Learning) were specifically concerned to measure pronunciation quality before and after some training sessions with the CAPT system. The results of these prototypes reported statistically significant higher production post-test differences of participants in the experimental group (see Tables 7.19 and 7.28).

A high consistency between metrics of objective and subjective pronunciation improvement has been observed (*Issue 2.1*). In particular, pronunciation improvement rate values in the prototypes of the Guided Learning experiment were assessed not only by human raters, but also by ASR systems and an objective game score (G) which highly correlated with the subjective scores provided by the raters regarding the pre/post-tests (see Tables 7.21, 7.26, 7.28). This score could be used to illustrate learner's interaction results with the CAPT system and also to recommend personalized activities to the users, saving time and human costs and resources, since it can be used to assess a large amount of students.

A measurable pronunciation improvement has been reported in each one of the experiments (*Issue 2.2*). From a quantitative point of view, the pronunciation improvement reported in the results of both competitions (TipTopTalk! and COP prototypes) showed better scores at the end of the experiment than at the beginning. These improvements were higher in the explicit competition, COP, being the differences statistically significant in both types of activities (see Table 7.37). In the case of the prototypes of the Guided Learning experiment (English Vowels and Japañol), the results obtained from the students of the CAPT system group (experimental) were compared to the in-classroom groups. This question was positively settled since a statistically significant higher pronunciation improvement was reported in CAPT-condition student's (see Tables 7.19 and 7.28). However, it must be pointed out that it was not questioned the efficacy of the instructor skills nor the in-classroom methodology.

The real difficulties of users (production tasks as the most difficult type of activity and the most difficult phonemes in the activities of the experiments) were also revealed in the results of the experimentation (*Issue 2.3*). Regarding the TipTopTalk! and COP prototypes, the success rate values of discrimination and production activities indicated that users found more difficulties with the latter type of activities, in both Training and Playing modes, since learners needed more time and achieved lower success rates in production activities (see Figure 7.10 and Table 7.11, and Table 7.34, respectively). This result agrees with the current trend in SLA [36]. Regarding the mode of activity, success rate values displayed in Table 7.11 and Table 7.34 of the TipTopTalk! and COP prototypes, respectively, reported better Playing mode rates than the Training mode ones. This could be due to the fact that students selected the most difficult sounds activities in the Training mode to improve their performance in the Playing activities as explained by themselves in the focus group sessions ("*I trained words that I did not uttered correctly*", "*I only trained when there were words I did not understand well*", "*I trained because I could listen to the words I found impossible to pronounce correctly*", "*I liked to be able to choose specific phonemes to practice in the Training mode*", "*I trained a little, only when I had problems with some words*", see Section 7.5.7). Furthermore, results derived from the English Vowels and Japañol prototypes reinforced the same idea of being production activities more difficult than discrimination ones. Learners needed more attempts and achieved lower success rate values in production activities than in the rest. Participants were usually better at discrimination than production in both prototypes (7.2 vs. 12.6 tries in Table 7.14 of the English Vowels prototype and 8.5 vs. 9.8 tries in Table 7.22 of the Japañol prototype). Finally, in this two prototypes it was also analyzed each contrasting lesson's phonemes and sounds, indicating which one was the easiest and most difficult (see Tables 7.15 and 7.16 of the English Vowels prototype, and Tables 7.23, 7.24, and 7.25 of the Japañol prototype).

It was, therefore, important to limit potential sources of frustration in the interaction with the CAPT systems (type of activities, Playing or Training modes and specific phonemes/sounds, *Issue 2.3*). The ASR technology used in production events (the most difficult activities) could have been a source of such frustration [13], [65] since it was not guaranteed that a correct learner's utterance is always recognized (see Section 8.1.1). Comments gathered from the focus group sessions reinforced this idea ("*I felt frustrated after failing consecutively*", "*I felt frustrated when the timer continued counting while the ASR system was evaluating my utterance*", "*I felt frustrated sometimes because the CAPT tool did not accept several words*", see Sections 7.2.6 and 7.5.7). To provide a partial answer to this problem o to these problems, we restricted the maximum number of production attempts per word and included personalized feedback techniques for this reason. Another source of frustration was the challenging policy carried out in the COP prototype. Users got frustrated because some players did not answer to the challenges submitted, as reported in the focus group sessions ("*There were people who did not accept my challenges and I could not go forward in the competition*", "*The first challenges of the competition were a disaster because so much challenges were ignored*", "*When I saw myself in low positions of the leaderboard, I quit playing*", "*I felt I submitted so much challenges and so many were unanswered*", see Section 7.5.7). Limiting the number of daily challenges and hiding inactive players were some of the solutions adopted to overcome this problem.

8.2.1 Limitations

Specific isolated studies of the several factors which affect CAPT's functioning and success as a whole are still needed. The efficiency of the proposed CAPT systems is attributable not only to the intensity of training, but also to its elements as a whole. That is, the training activities involved, the feedback and assessment provided, the selection of minimal pairs, the use of TTS and ASR systems, and the CAPT design that connects all of them into an educational tool. Although they are mainly responsible for the reported success of the CAPT systems, it cannot be stated the specific role they have taken individually in the learning process since they have not been tested in isolation.

Further analysis and experimentation on the generalization of our approach to other phonemes, words, languages and long-term persistence of learning should be on their way in the near future. For instance, the post-tests of the Guided Learning and Competitive Learning experiments were performed one week after the training sessions. A delayed post-test would be required to check learning retention. However, it is difficult to verify the effect of the pronunciation improvement amelioration after a long period of time since the lifetime of a thesis is short and it is very complicated to find real participants who agree to collaborate continuously.

A closer comparative research between the activity and results of the CAPT system and those of human-led instruction (video-taping or log monitoring) in the Guided Learning experiment might have helped to obtain a more detailed knowledge on the possibilities of CAPT. The focus of the research by no means included a detailed analysis (assessment or consideration) of the particular instructor's way of teaching. It was intended to ascertain whether the CAPT system obtained significant/acceptable results as a teaching tool. In fact, the results obtained by the system have proved that are comparable to those obtained by a particular in-classroom procedure, carried out by an experienced instructor, and sanctioned by a respectable

institution. Within this particular experimental frame, and the ethics involved, it was not elaborated further comparisons or conclusions.

Objective quality metrics of ASR systems must be taken with precaution since it is not always known what they really measure. Even though the objective assessment with the game score provided in this thesis can save time and human resources, it will not be as accurate as possible as subjective rater assessment. It should be taken with precaution and as a complementary resource. Besides, the *g*–score value must be carefully adopted since it must be necessary to know how it is calculated (i.e., nativeness-like, intelligibility, or isolated phonemes of the utterance). In addition to the position of the expected utterance in the list of hypotheses provided by the ASR system, both parameters should be adjusted to the purpose of the experiment. Also limiting the number of attempts per activity and offering specific feedback techniques to avoid possible learner's frustration and fatigue.

The focus of this thesis has been the pronunciation improvement at the segmental level. However, its combination with the suprasegmental level must be also considered for acquiring a complete pronunciation improvement competence [33]. In fact, the realistic goal sanctioned by most scholars in the field is none other than recognizability, usually expressed as *intelligibility* [5], [59]. Transferred productions are also not free from controversy, and one might claim, with the support of most experts today, that native-like accuracy is not a realistic goal in pronunciation teaching. Although there is a long tradition of organizing pronunciation teaching around an inventory of phonemes (segmental), some scholars, often considering that native-like pronunciation of segments is rarely attained by L2 learners, have proposed the suprasegmental level (prosody, intonation, speed, fluency, etc.) as the proper target of pronunciation teaching. While the segmental dimension has been historically favored, most pronunciation training programs today try to strike a fair balance between both levels. The CAPT systems proposed in this thesis aim at complementing pronunciation training at the segmental level. However, the suprasegmental level should also be considered and studied if it is ever necessary to attain a full understanding of the potential of CAPT.

8.3 Game-based Learning with CAPT Systems

The game elements included in the prototypes of the experimentation of this thesis have proved beneficial when integrated into CAPT systems in order to motivate some learners to keep on playing by their own (RQ3). As mentioned in the state-of-the-art Chapter 3, to date there is no empirical study about a learning competition for L2 pronunciation training with a social CAPT system. However, the analysis of the effects of the design of gamification elements on student's motivation is receiving an increasing interest in the literature. In LL there are still few empirical studies that have addressed this issue, and specially, in pronunciation training with CAPT systems (see more details in Section 5.2). In that sense, some have focused on the study of the differential effects of gamification elements on the intrinsic motivation, competence, and performance [153], [159]. Meanwhile, other studies have addressed the effects of explicit competition on student's behavior and motivation [143], [201]. In particular, several gamification elements have been included in the three experiments focused on game-based learning (Minimal Pairs, TipTopTalk!, and COP prototypes). On the one hand, the Minimal Pairs prototype served as a first

approximation, being a short single-session test. In this experiment, it was only included a scoring, a timer, a counter of right and wrong attempts (scoreboard), and visual animations and sounds when the users interact with the CAPT system. It was also gathered several opinions from the users about game improvements for future experiments in the focus groups session (see Section 7.2.6). On the other hand, the following two prototypes (TipTopTalk! and COP) integrated more gamification elements into two levels of competition, implicit and explicit (without or with challenges, respectively), during a one-month competition protocol (see Sections 7.3 and 7.5, respectively). Particularly, they included points, leaderboards, scoreboards, badges, prizes, performance graphs, avatars, progress visualization, limited number of attempts, sounds, visual animations, and the interaction (implicit or explicit) with other players.

Results obtained from the social experiments of this thesis led to find out some differential effects on student's game motivation with the CAPT system (RQ3). The theory states that, in addition to increase learner's motivation, the main advantage in using a gamification design strategy consists in the possibility of providing individualized and comprehensive feedback while keeping users comfortable and active to progress at their own pace in an anxiety-free context. Regarding motivation, the TipTopTalk! prototype's competition results about the answers to the final questionnaire showed a positive predisposition to the activities of the game (a 75% of answers reported that learners found very fun the game dynamics, and a 95% of them found very easy to understand these dynamics; see Section 7.3.7). Besides, although there was not an explicit interaction with other players neither any kind of pressure nor external motivation of the research group, learners were interacting with the CAPT system individually on average almost four days and performed 240 discrimination and/or production events every day (see Figure 7.8 and Table 7.10, respectively). However, the average participant's activity was quite irregular with sudden high peaks some days and low activity in the rest of the days (see Figure 7.9). In the case of the COP prototype, results showed that it had also a positive effect on the most active student's motivation. First, it was found a constant activity in the most active students in the explicit competition condition (see Figure 7.24). These time and intensity results of gaming activity agree with other studies, such as [138], which found a positive effect of competition on player's motivation. Unlike those previous studies which stated that competition and gamification elements, such as points, leaderboards, among others, could diminish the intrinsic motivation and engagement [128], [202], results of the experiments presented in this thesis showed a positive effect on most active student's motivation, in agreement with other previous studies that considered competition as a motivational trigger stimulating engagement and persistence [203], [204], [205].

Results also led to find out some differential effects on student's game performance with the CAPT system (RQ3). Participants of the TipTopTalk! prototype achieved, on average, a positive discrimination and production improvement of 9% and 1%, respectively, as reported in Figure 7.10. However, in general, and more specifically in the case of most active players, a habituation factor led to a fall in motivation and performance after protracted use. On the other hand, according to the results gathered in the COP prototype, the challenge-based competition had positive effects on student's production and discrimination success rates. Table 7.37 showed in this case statistically significant differences between the results at the beginning and at the end of the competition in both discrimination and production activities, being higher for the most active players.

Including challenges in the CAPT system of the COP prototype had a positive influence on the student's performance of the Constant (see Section 7.5.6) group (RQ3). In order to study in greater depth the effects of a challenge-based competition on user's performance with the CAPT system proposed in this thesis (COP prototype), it was carried out a comparison between the COP prototype participants, categorizing them in three statistical tertiles according to their activity in the competition (number of matches with participation): Constant, Habitual, and Casual groups, from the highest to the lowest participation in challenges, respectively. Results showed that the most active and motivated participants in terms of game activity (Constant group) achieved the highest success rates in both production and discrimination activities in the Playing mode, with significant differences in comparison to Habitual and Casual groups (see Table 7.34). Comparing the Habitual and Casual groups, it was only found a significantly better performance in discrimination activities in the Playing mode by the Habitual group. Regarding performance in the Training mode, results showed that, in both types of activity, production and discrimination, the Constant group was better than the Casual one, and the Habitual group was better than the Casual one, but it was not find any statistically significant difference between the Constant and Habitual groups. These results support the statement reported in [138], which concluded that the perception of video game competitiveness has a strong effect on user's flow experience and satisfaction, due to the positive effect that overcoming challenges has on self-concept and in the sense of competence. Apparently, it might be that being more focused, putting more effort into the activity, and pursuing a goal, such as winning other players in challenges, improves the performance and proficiency in the case of the most active users (Constant group). However, the challenged-based competition could have decreased motivation in users with low possibilities to obtain the final prize (Habitual and Casual players).

Including challenges in the CAPT system of the COP prototype had a positive influence on the Constant student's motivation (RQ3). Despite the fact that some previous studies in educational games have pointed to the use of competitive conditions and the use of external rewards, such as points, leaderboards, and prizes, among others, to have a negative effect on the intrinsic motivation of students [128], [202], the results of this thesis do not agree with this statement. The answer to the question Q1.3 of the COP post-test questionnaire (Table 7.38) showed that, in the three groups, more than 75% of the participants selected the option "*improving my English pronunciation*" as the main incentive to use the game, and the analysis carried out did not show any statistically significant difference between groups. In that sense, despite the low disposition toward competition in some of the students, it must be pointed out that the introduction of controlled elements of competitiveness, such as a limited number of daily challenges, the restriction of being able to challenge only to the ten players above and below on the leaderboard and the clear and defined rules [123], is a motivational trigger for some students (Constant group), since the challenges they had to face require students to improve their skills to win [135]. The second main reason declared by users was climbing up the leaderboard (Q1.6). In particular, it was declared by more than the 90% of the Constant group, being this value statistically significant reduced in the other two groups (69.1% and 50.0%, respectively), reaching an average of 70.3% of positive answers. Therefore, as [138] stated, competitiveness can contribute to experience flow and increasing the sense of competence when the challenges are overcome. Results of the post-questionnaire carried out in the previous second experiment already anticipated the

positive predisposition of the users to challenge other players in the game (90%, fourth question of Figure 7.12).

Also, the challenges had a positive influence on the Constant student's proficiency improvement (RQ3). Some studies, such as [143], reported that competition has fewer positive effects on learning than non-competitive games. Specifically, in their study, competition made students less motivated and engaged with the additional learning materials provided in the game. Others, such as [140], conducted an experiment to compare the collaborative condition and the competitive one, reporting that their results did not show any difference between these two independent conditions. In order to gain greater depth into the effect of competition in learning outcomes in this thesis, it was analyzed if there had been any improvement in the pronunciation and discrimination skills. As opposed to [143], results presented in Table 7.37 showed that COP players had a significant improvement in both production and discrimination activities. According to the differences in the amount of improvement on learning among the three competition groups, results showed that, in both types of activities, the Constant group had a significantly higher improvement than the other ones. Besides, in production activities, the Habitual group had a significantly higher improvement than the Casual one. Judging from collected data, results showed that competitiveness and well defined situations with clear goals and feedback, could be useful elements in some types of learning activities of CAPT systems. This supports the idea of the most active players (Constant group) achieved better learning outcomes.

Less than half of the participants finished the data collection in the COP competition. Although a high abandonment rate has been reported in the COP competition (52% of the registered participants did not finish data gathering), this value was in tune with the minimum values in the current trend in online learning courses, since studies report 40% to 80% drop-out rates of online university classes [206], [207]. Besides, nowadays just 40% of video game players continue playing after the first day [208]. Also, studies report that learning games are at a disadvantage in comparison to pure entertainment games [209]. Particularly, in COP, 22 players abandoned after their first day session (12.1% of the total of 182 dropouts). The abandonment reasons were tried to be clarified by asking these users about the possible causes (Table 7.40). Most answers were technical reasons (42.4%, Q3.1) and lack of time (41.7%, Q3.2). The rest of main responses related to the likeability of the competition did not reach the 28% (Q3.3, Q3.4, and Q3.5). Lack of time can be associated to the great quantity of challenges needed to be constant in the game and to have the potential to achieve the rewards. Likeability answers were also understandable since not all players found the competition suitable for learning, and not all players were equally skilled for competing. However, in the latter case, there were 45 participants of the 182 who early abandoned who had an initial production success rate > 51%, 35 players achieved values over 70%, and 10 achieved at least a 87% initial production success rate. In comparison to the 30 most active players of the Constant group (56 players), they achieved minimum, average, and maximum rates of 51%, 70%, and 87%, respectively.

8.3.1 Limitations

Further experimentation on the ways to increase user engagement would be necessary, given the high rate of abandonment in the COP social competition. From the analysis of the reasons for early abandonment reported by the participants,

we concluded that peer reaction to send challenges is still to be rethought in order to avoid discouragement in the most active players. As a future work it might be interesting to offer a wider range of rewards to motivate to keep on training and playing not only the most active users, but also the rest of the players.

Final rewards of the competitions could have been influenced the results. The final compensation provided to the participants of the experiments, such as diplomas, academic certifications, and rewards depended directly, at that specific moment, on the financial support by the research project of the experiment carried out. This variable could have influenced the extrinsic motivation of the participants [210].

Finally, all these aspects could limit the extent of the generalization of the results. More research is necessary to explore the role of different personal variables, such as dispositional competitive personality traits, skill levels in the activities, and previous levels of motivation for learning, among others. It is also needed more research that explores the different levels and conditions of the competition with respect to the type of learning activities, such as a collaborative CAPT system or a control group with only training activities. Furthermore, it would be necessary to explore other potential negative risks of the use of digital games on learning contexts, such as their addictive potential or their possible negative effects on social relations in the classroom.

8.4 Summary

In this chapter the research objectives have been validated and the research questions established at the beginning of this thesis have been answered, since the research methodology followed has been justified and evaluated critically. Furthermore, the most relevant limitations derived from the experimentation have been pointed out.

Firstly, the research question related to the inclusion of state-of-the-art TTS and ASR systems into CAPT projects has been positively settled (RQ1 and *Issue 1.1*). Although there are scarce empirical experiments in the literature that evaluate the effectiveness of TTS and ASR technology in CAPT systems, experimental results of this thesis evidenced a large number of events with these technologies by the users. Most of the learners who participated in the experimentation achieved a statistically significant pronunciation improvement after training with the proposed CAPT systems, proving these technologies can be intended to LL processes in a non-obstructive way under the supervision of experts. Even though there is still some controversy about the educational purpose of these systems, they have been proved to be a useful resource to assess users depending on their L2 level, as reported in each experiment of this thesis (*Issue 1.1*). Besides, the CAPT systems developed for the prototypes of the experimentation have been benefited from the possibilities that these technologies offer, such as pronunciation quality metrics and immediate feedback.

Secondly, the main results related to the training methodology proposed and their implications on user's pronunciation improvement have been summarized (RQ2, *Issue 2.1*, *Issue 2.2*, and *Issue 2.3*). These activities and minimal pairs lists have been defined and elaborated by experts with the help of a semi-automatic protocol designed during this thesis. On the one hand, a great quantity of training activities

performed by the learners during the experiments that led to a significant pronunciation improvement has been reported (*Issue 2.1*). On the other hand, measuring participant's pronunciation improvement has been possible not only with subjective raters' scores, but also with an objective score obtained automatically from the technology integrated (*Issue 2.2*). Finally, the real difficulties of the students during the experiments have also been possible to find, such as the most difficult type of activities, phonemes, and sounds (*Issue 2.3*).

The last research question of this thesis, RQ3, has also been answered positively. Several game elements have been included into two CAPT system's prototypes. Results have reported positive effects on some user's motivation and performance, both subjectively and objectively, with questionnaires, focus groups, and game logs, respectively. A novel contribution in which students can challenge each other via pronunciation activities has been discussed, in which the most active players showed high rates of motivation and have achieved significant pronunciation improvement.

Part IV

Conclusions

Chapter 9

Conclusions

In this PhD thesis, we have addressed the design and validation of an innovative CAPT tool for smart devices for the training and assessment of L2 pronunciation. It has been focused on segmental pronunciation with a specific set of methodological choices, such as L1–L2 connection, particular lists of minimal pairs, an exposure–perception–production training cycle, and the inclusion of ASR and TTS technologies in the CAPT system. This thesis has also been carried out within a multidisciplinary research framework and institutional collaborations, and two basic different versions of the CAPT system were designed and tested with real users to analyze and discuss the final outcomes in detail. The first one incorporated game elements to keep individuals motivated while training at free will. The second one removed the game elements and imposed a controlled training protocol. In the former case, users kept on training and were motivated while practicing anywhere. Experiments of the latter version were carried out in a laboratory, in a shorter and controlled training protocol. Results evidenced relative pronunciation improvement (production and perception skills) of the users who trained with the CAPT systems, being higher in the case of most active learners.

This dissertation has carried out a multidisciplinary approach, combining TTS and ASR technologies, gamification techniques, phonetics and pedagogical considerations, and software development of smartphone applications for L2 pronunciation and testing. Furthermore, the scope of this dissertation has covered real users from different locations and nationalities who participated in each experiment, overcoming some issues that it involves, such as searching, availability, enrollment, and rewards. Hence, an important contribution to the field of language learning and pronunciation tutoring systems through new technologies has been provided. As future work, the great quantity of data gathered from the experimentation will permit researchers not only to automatically characterize user's L2 speech, but also to improve, personalize, and increase the content of the CAPT system.

This chapter contains the conclusions of this thesis, including the main ideas, scope, and contributions to the state-of-the-art. Besides, some lines of future work are also discussed. Finally, a list of the main achievements and attributions obtained throughout the thesis is presented.

9.1 Conclusions

In this thesis, off-the-self ASR and TTS technologies have proved that can be effectively incorporated in a non-obstructive way to the L2 CAPT tools developed in this work. The results reported of perception and production activities mediated by speech technology, strongly correlated with the expected level of L2 of the user, being native speakers scores better than advanced and beginner non-native learners ones, respectively. These speech technology systems have been used in a non-obstructive and user-friendly way which has permitted to reuse their components in consecutive experimentation.

The incorporation of current speech technology, in terms of ASR and TTS, was able to provide a very useful and didactic instrument in the pedagogical version that can be used complementary with other forms of SLA. The off-the-self TTS and ASR systems included in the experimentation have proved to be particularly useful for increasing the amount of participation, guiding feedback and immediate diagnostic, and to provide model pronunciations for the learners. Thus, for any technological complement to be truly effective, it must be subordinated to carefully designed methodological frameworks that also include human interaction. It therefore represents an interesting attempt to exploit the affordances of technology which was not initially conceived of as educational, for the purpose of teaching pronunciation at the segmental level.

The methodological decisions implemented in the different version of CAPT tools designed and validated in this work allowed to reliably measure the relative pronunciation improvement of the individuals who trained with them. The training methodology was partially grounded on a well-known and recognized methodological learning approach (NCM), and focused on specific phonemes of the target L2 language that are difficult for student's L1. Three main new methodological aspects of the CAPT systems have been proposed (which usually CAPT tools in the literature miss to include). First, the incorporation of speech production and perception technology and techniques. Second, a collection of thoroughly designed and performed tests prior and after treatment. Finally, a set of well designed selection of training modes, activities, and phonetic phenomena from a teaching and learning perspective. In particular, in contrast to most experiments reported in the literature, which include isolated exercises, in this thesis we have proposed a new extended methodology based on the combination of different activities offered by a system with ASR and TTS systems, depending on the results obtained by the user with the aim of training and improving L2 pronunciation.

An automatic data gathering method for CAPT results which allowed to provide specific feedback to users automatically and to analyze all results at the end of the experiments has been carried out; whereas in similar experiments in the literature users are required to write down their own results and advance to the next exercise manually or record their own interactions for future processing.

The minimal pair lists included in the CAPT tools developed were elaborated with a novel semi-automatic protocol, taking into account learner's L1 and L2 and the specific known limitations of ASR and TTS technologies. Modern TTS systems can be seriously considered by developers of CAPT systems who need to generate pronunciation models for isolated words; whereas the use of ASR systems as part of CAPT tools with isolated words requires a careful pre-selection of language elements, such as sounds, words, and contexts to be included in the exercises. Working

with minimal pairs and such speech technologies also led to take some considerations about homophones, false alarms, infrequent words, and out-of-context words.

Throughout the thesis experimentation, some of the techniques described in the literature to provide an innovative, adaptive, and clear feedback in the CAPT systems have been integrated in our prototypes. **In particular, different corrective feedback techniques in which users were guided automatically, depending on the individual's results, to overcome the proposed training, proved to be useful and effective in the developed CAPT systems.** To the best of our knowledge, there is not a CAPT system that guides users through a personalized and carefully designed path of activities to achieve better pronunciation skills based on their results. For instance, users with a low performance are redirected to specific and easier exercises that help to achieve better results. In particular, and as a novelty, after a determined number of consecutive failures, the system executes an explicit corrective feedback response that invites users to listen to the synthesized version of the problematic word. In this case, the TTS output did not interrupt the natural process of acquisition of sounds while training. On the other hand, users with a good performance needed a smaller number of training activities to reach the final objective.

Positive results from both subjective and objective assessment techniques for pronunciation improvement during and after using the CAPT systems developed in this thesis have been reported. The n–best list of string candidates and the scores provided by the ASR system integrated into the CAPT tool have been used to assess user's utterances and correctly classify them by different levels of pronunciation. In addition to the binary score obtained in discrimination tasks, this information was also shown to the user as feedback of the current training task and mode score. The strong correlation between the subjective scores in the pre-test and post-test and the ASR ones of the English Vowels and Japañol prototypes evidenced that the ASR technology is adequate for pronunciation assessment.

Results reported from the prototypes of the Guided Learning experiment are very promising since the students who worked with the CAPT systems achieved better pronunciation improvement values than their peers in the traditional in-classroom instruction group. Comparison of pre/post-test results explicitly showed the usefulness of the tool as a supporting instrument to normal L2 in-classroom lessons. The pace of during the training process showed the effectiveness of the mandatory and careful design of the exposure–discrimination–production cycle, the corrective feedback, the well-informed teaching approach and the "satisfactory quality" of the TTS and ASR. As a contribution of this thesis, right and wrong attempts of perception and production activities during training with the CAPT system have helped to elaborate a final CAPT final score per student, G. This score strongly correlated with the final subjective score provided by the raters to each learner in the post-test, so as it can serve as an useful indicator of training benefits and helped, not only to save human and costs resources, but also to adapt and recommend exercises during the training for future experiments. In the two competitions carried out in this thesis (TipTopTalk! and COP prototypes) an average score of the successful production and perception attempts during the same time window at different periods of the experiment has been reported. In both experiments better scores were reached at the end of the competitions in both activities (except for the advanced level proficiency users in the production activities TipTopTalk! competition), being statistically significant in the COP competition. Results reported from the COP prototype showed that the most active players achieved better pronunciation results.

Finally, game elements had a positive influence on user's motivation, performance, and learning in the different CAPT systems developed in this thesis. This is an important result given the increasing interest of educators to discover new ways to motivate students and to encourage effective uptake. The robust and scalable architecture of the games, scoring system, and metrics have also contributed. Particularly, a set of gamification elements has been included into the CAPT-based learning games developed for keeping users motivated while they are training their L2 pronunciation. First, and individualistic approach in which users play with the system in an implicit competition with themselves (TipTopTalk!). In particular, a learning game has been developed in which users can select the activities to perform at their own will. Second, and also as a contribution of the thesis, a challenge-based game for pronunciation training, COP, has been carried out under the guidelines of a learning competition [123]. **This competition proved to be a positive motivational factor, specially for the most active users, whose intensive use of the game allowed them to achieve significant L2 pronunciation improvement along time.**

9.2 Future Directions

Although the results of this thesis are satisfactory, there are some aspects that can be improved and give way to new lines of future work. In particular, the experimental framework followed in this thesis (see Figure 7.1) leads to **three main lines of future work**: **(1) a systematic study of L2 speech automatic characterization of the speech data gathered during the experimentation; (2) a more personalized and individualized adaptation of the CAPT system to the learner; and (3) a predictive behavior modeling of user's interaction with the CAPT system.** The systematic study is considered as a fundamental aspect since corrective feedback is a basic pillar of pronunciation instruction. This step was briefly started at the end of this thesis when analyzing with Kaldi the audio dataset gathered from the participants of the University of Seisen of the Japañol prototype. **Future work will consist on analyzing and designing specific speech recognition algorithms for the identification of pronunciation mistakes associated with key features of proficiency level characterization.** In particular, it would be useful to be able to determine the set of key features when correlating pronunciation mistakes with assessment grades of experts and, on the other hand, to facilitate the extraction of this set from automatic classification systems that can guide the formulation of personalized recommendations on place and manner of articulation.

Finding new techniques for adapting a CAPT system to the user in a more personalized and individualized way would also help to improve not only her/his pronunciation improvement outcomes, but also her/his motivation degree during the pronunciation training. That is, keeping the motivation to play and therefore to learn. In this thesis the first steps for the resolution of this task have already been carried out. However, most of them have consisted on suggesting several generic activities and advises of the most common L2 mistakes previously designed by experts. Applying new techniques would allow the CAPT system to determine in real-time time which path of activities to follow in order to promote a wider acquisition of the total number of sounds, and which type of corrective feedback to provide for each user, individually, due to the possibility of analyzing the data gathered while training. Feedback would also provide qualitative and quantitative information on the deviation of the learner's production from a range of acceptable pronunciations.

An analysis of the relationship between the CAPT system's design, the strategy followed by the learners, and their final outcomes could be useful for categorizing and predicting user's behavior with the system. This learner modeling could predict and improve student's performance at intermediate training stages; whereas the rate of early abandonment could be also diminished. In fact, at the end of the period of this thesis, the PhD candidate has undertaken an international research stay in a prestigious research center in educational innovation in order to deepen this topic with the data gathered in the COP prototype (see Section 9.3.5 for more details). A new and promising collaboration in the field of learning analytics for learning design has started and will continue after the end of this predoctoral period.

Furthermore, several degree final (academic) projects and international/national collaborations are currently maintained with undergraduates, language research centers, experts, and instructors. On the one hand, the testing phase of the Estoñol prototype is being carried out at the University of Tartu, Estonia. On the other hand, the data gathered with the last batch of speakers of the University of Seisen, Japan (Japañol prototype) is also being analyzed in collaboration with the *Servei de Tractament de la Parla i del So* of the Autonomous University of Barcelona.

Finally, in addition to the direct results obtained during the experimentation, two speech corpus datasets have been automatically gathered as part of the training activities, which can be used and extended in future experiments. Even though the size of the corpora data analyzed in each experiment of this thesis has been sufficient to obtain statistically significant results and to publish them in several journals and conferences, it is necessary to continue gathering audio samples from more speakers, not only from the adult university area (as it has been principally during the thesis), but also from other populations, such as children or teenagers, who are more used to technology. Also, the more quantity of speech utterances, the better quality of the specific-purpose Kaldi ASR system that is being developed. This increase in the corpus size might also facilitate the exploration of new strategies for automatic L2 speech classification, both at the segmental and suprasegmental level, not only by developing a personalized and specific-purpose ASR system, but also by applying DNN algorithms with Kaldi, which need a large number of data to obtain reliable results.

9.3 Achievements and Attributions

The list of contributions to international journals and conferences related to this thesis is presented below.

9.3.1 Journal Publications

Journals indexed in the Journal Citation Report (JCR):

1. **JCR Q2**: 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 Transactions on Learning Technologies*, vol. 13, no. 2, Mar. 2020. DOI: 10.1109/TLT.2020.2980261 [23]

2. **JCR Q1**: 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. 74250–74266, Apr. 2020. DOI: 10.1109/ACCESS.2020.2988406 [31]

Other journals:

3. 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", *Journal of Estonian and Finno-Ugric Linguistics (ESUKA – JEFUL)*, vol. 10, no. 1, pp. 89–104, Nov. 2019. DOI: 10.12697/jeful.2019.10.1.05 [30]

4. C. Tejedor-García and D. Escudero-Mancebo, Üso 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. 10, no. 1, pp. 355–363, Jan. 2018. [29]

9.3.2 Conference Publications

1. 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 Proceedings of *IberSpeech 2018*, Barcelona, Spain, Nov. 21–23, 2018, pp. 97–101. DOI: 10.21437/IberSPEECH.2018-21. [25]

2. C. Tejedor-García, V. Cardeñoso-Payo, and D. Escudero-Mancebo, "Japañol: a mobile application to help improving Spanish pronunciation by Japanese native speakers", in Proceedings of *IberSpeech 2018*, Barcelona, Spain, Nov. 21–23, 2018, pp. 157–158. [26]

3. C. Tejedor-García, "Design and evaluation of a mobile application for second language pronunciation training based on minimal pairs", *in XXXIV Congreso Internacional De La Sociedad Española Para El Procesamiento Del Lenguaje Natural (SEPLN)*, Seville, Spain, Sep. 21–23, 2018. pp. 7–11. [28]

4. T. Kimura, C. Tejedor-García, M. J. Machuca, A. Ríos, and D. Escudero-Mancebo, "Japañol, a Computer Assisted Pronunciation Tool for Japanese Students of Spanish Based on Minimal Pairs", *Abstracts of 2nd International Symposium on Applied Phonetics*, Aizu, Japan, Sep. 21–23, 2018. [27]

5. 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 SLaTE, Stockholm, Sweden, Aug. 25–26, 2017, pp. 26–30. DOI: 10.21437/SLaTE.2017-5. [24]

6. C. Tejedor-García, D. Escudero-Mancebo, C. González-Ferreras, E. Cámara-Arenas, and V. Cardeñoso-Payo, "TipTopTalk! Mobile application for speech training using minimal pairs and gamification", in Proceedings of *IberSPEECH 2016*, Lisbon, Portugal, Nov. 23–25, 2016, pp. 425–432. [19]

7. 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 Proceedings of *IberSPEECH 2016*, Lisbon, Portugal, Nov. 23–25, 2016, pp. 177–186. [20]

8. 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 Proceedings of *Interspeech*, San Francisco, SF, USA, Sep. 8–12, 2016, pp. 1178–1179. [18]

9. 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," *Abstracts of New Sounds Aarhus, 8th International Conference on Second Language Speech*, Aarhus, Denmark, 2016, pp. 160. [22]

10. 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 Proceedings of *IFCASL*, ser. Workshop on "Feedback in Pronunciation Training", Saarland, Germany, Nov. 5–6, 2015. [21]

11. 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 Proceedings of *SLaTE 2015*, Leipzig, Germany, Sep. 4–5, 2015, pp. 125–130. [17]

9.3.3 Attendances and Participation in Conferences and Workshops

1. *International Conference*. "LASI Nordic 2019" Nordic Learning Analytics Summer Institute. Tallinn, Estonia. 28–30 Aug. 2019. [Online] Available: https://lasi2019.tlu.ee/. Accessed on: Nov. 3, 2019

2. *International Conference*. IberSPEECH 2018. X Jornadas en Tecnologías del Habla and the V Iberian SLTech Workshop events. Barcelona, Spain. 21–23 Nov. 2018. [Online] Available: http://iberspeech2018.talp.cat. Accessed on: Nov. 3, 2019

3. *International Conference*. SEPLN 2018. XXXIV Congreso Internacional De La Sociedad Española Para El Procesamiento Del Lenguaje Natural. Sevilla, Spain. 19–21 Sep. 2018. [Online] Available: http://www.sepln.org/en/headlines/news/34th-international-conference-sepln. Accessed on: Nov. 3, 2019

4. *International Conference*. SLaTE 2017. Speech and Language Technology in Education. Stockholm, Sweden. 25–26 Aug. 2017. [Online] Available: https://www.isca-speech.org/archive/SLaTE_2017/. Accessed on: Nov. 3, 2019
5. *International Conference*. LII Congreso internacional de la *Asociación Europea de Profesores de Español* — Universidad de las Palmas de Gran Canaria. Las Palmas de Gran Canaria, Las Palmas, Spain. 24–28 Jul. 2017. [Online] Available: http://www.aepe.eu/actividades/congresos/. Accessed on: Nov. 3, 2019
6. *International Conference*. IberSPEECH 2016. IX Jornadas en Tecnologías del Habla and the V Iberian SLTech Workshop events. INESC-ID Lisbon, RTTH and SIG-IL. Lisbon, Portugal. 23–25 Nov. 2016. [Online]. Available: https://iberspeech2016.inesc-id.pt/. Accessed on: Nov. 3, 2019
7. *International Conference*. The "LASI España 2016" International workshop - University of Deusto (UD) and SNOLA (Spanish Network of Learning Analytics). Deusto Campus, Bilbao, Spain. 27–28 Jun. 2016. [Online] Available: http://lasi16.snola.es/. Accessed on: Nov. 3, 2019
8. *National Conference*. II Jornadas de Investigación en Tecnologías de la Información y de las Comunicaciones (University of Valladolid). Valladolid, Spain. 28 Apr. 2016. [Online] Available: https://goo.gl/AQ3h6V. Accessed on: Nov. 3, 2019
9. *International Conference*. IFCASL Project 2015. Workshop on "Feedback in Pronunciation Training". Saarbrucken, Germany. 5–6 Nov. 2015. [Online]. Available: http://www.ifcasl.org/index.html. Accessed on: Nov. 3, 2019

9.3.4 Intellectual Property Register

The second prototype of the Non-guided Learning experiment, TipTopTalk!, has been legally registered under the intellectual property protection of computer software (May 20, 2016): *TipTopTalk! Aplicación móvil para la mejora de la pronunciación multilingüe mediante la utilización de pares mínimos y gramificación* (Central Registry of Intellectual Property, Ministerio de Educación, Cultura y Deporte, Spain. Request number: VA–170–2016, registration number: 00/2016/2525).

9.3.5 Research Stays

1. **National stay (one month):**
   - **Start date:** 2017/06/19 (second year of the PhD dissertation)
   - **End date:** 2017/07/23
   - **Financing entity:** Ministerio de Economía y Competitividad (Spain), project key: TIN2014-59852-R
   - **Host institution:** *Servei de Tractament de la Parla i del So* (department of Spanish Philology, Facultat de Filosofia i Lletres, Autonomous University of Barcelona), Barcelona, Spain
   - **Summary.** In the last decade we have witnessed the commercial success of foreign language applications that incorporates speech synthesis and recognition, based on different pronunciation improvement techniques. During the first two years of his thesis research, the PhD candidate has

been experimented with various applications and techniques that have been positive for the improvement and evaluation of foreign speech. The objective of this stay is to focus on the pronunciation of the Spanish language as L2 for native Japanese speakers. This is a pair of L1 and L2 which has not been worked with until now in the thesis. This is a multidisciplinary research project, in which there is a collaboration with the department of Philology Spanish from the Autonomous University of Barcelona. It is necessary their expertise in the field to provide the pedagogical data to the Japañol prototype. Once the investigation is finished it will be proposed to maintain contact to analyze and publish the results derived after a potential experimental testing campaign with the system developed.

2. **International stay (three months)**:

   - **Start date:** 2019/09/26 (last year of the PhD dissertation)
   - **End date:** 2019/11/26
   - **Financing entity:** University of Valladolid: Predoctoral short-term fellowship — 2019 (*Movilidad doctorandos ayudas para estancias breves en el desarrollo de tesis doctorales — Convocatoria 2019*)
   - **Host institution:** Centre of Excellence in Educational Innovation, University of Tallinn, Tallinn, Estonia
   - **Summary**. The main objective is to improve the analysis of the available game results of a CAPT learning game, COP, developed in one of the experimental prototypes of this thesis. This game establishes a competitive protocol to improve English pronunciation as a foreign language. COP has been designed to facilitate competitive mechanisms among players, launching and accepting challenges and obtaining extra points or penalties depending on the results. A preliminary pronunciation improvement analysis has already been carried out, but much more information is available that can provide relevant data on the different player profiles, the game strategies adopted and the relationship of all this with the pronunciation improvement. In addition to explore new techniques of data analysis and time series to obtain results on player's behavior and their relationship with learning, it is also expected to obtain relevant information on those aspects of the game protocol that would need to be modified to improve future versions of the game.

9.3.6 Speech Datasets

As a result of the automatic gathering of information, a significant amount of speech data from the last two prototypes of the experimental procedure has been recollected (see Appendix D) since Google enabled the possibility of keeping the audio utterances when using the GCSTT online service (as explained in Section 6.4.1). This technique must be taken into account for future experimentation as an effective and fast way of recollecting decentralized speech data for elaborating datasets.

9.3.7 Software Resources

As a consequence of the intensive software development carried out (see details in Section 6.4.3), several software applications have been released during the thesis research which are freely available:

1. ***Minimal Pairs* Android application:**
   - Experiment: Alpha (see details in Section 7.2)
   - Prototype: Minimal Pairs
   - URL: https://eca-simm.uva.es/es/proyectos/capt/minimal-pairs/
2. ***TipTopTalk!* Android application:**
   - Experiment: Non-guided Learning (see details in Section 7.3)
   - Prototype: TipTopTalk!
   - URL: https://eca-simm.uva.es/es/proyectos/capt/tiptoptalk/
3. ***English Vowels for Spanish People* Android application:**
   - Experiment: Guided Learning (see details in Section 7.4)
   - Prototype: English Vowels
   - URL: https://eca-simm.uva.es/es/proyectos/capt/english-vowels-for-spanish-people/
4. ***Japañol* Android application:**
   - Experiment: Guided Learning (see details in Section 7.4)
   - Prototype: Japañol
   - URL: https://eca-simm.uva.es/es/proyectos/capt/japanol/
5. ***Estoñol* Android application:**
   - Experiment: Guided Learning
   - Prototype: Estoñol
   - URL: https://eca-simm.uva.es/es/proyectos/capt/estonol/
6. ***Clash of Pronunciations* Android and web application:**
   - Experiment: Competitive Learning (see details in Section 7.5)
   - Prototype: COP
   - URL: https://eca-simm.uva.es/es/proyectos/capt/clash-of-pronunciation/

9.3.8 Attributions

1. **Best demo award** at IberSPEECH'2016 international conference, Lisbon, Portugal (Nov. 25, 2016): *TipTopTalk! Mobile application for speech training using minimal pairs and gamification* —C.Tejedor-García, D. Escudero-Mancebo, C. González-Ferreras, E. Cámara-Arenas and V. Cardeñoso-Payo.
   URL: https://iberspeech2016.inesc-id.pt/index.php/award-winners/
2. **Best poster award** at II Jornadas de Investigación en tecnologías de la información y de las comunicaciones, University of Valladolid, Valladolid, Spain (Apr. 28, 2016).
   URL: https://goo.gl/AQ3h6V
3. **Prometeo award**: Plan TCUE 2015-2017, General Foundation of University of Valladolid, Valladolid, Spain (Jun. 6, 2015).
   URL: https://funge.uva.es/innovacion/tcue/programa-prometeo/

9.3.9 Funding

The research carried out in this PhD thesis has been mediated by the context of the following research projects and research fellowships:

1. Title: Social video games for training and improving L2–Spanish pronunciation (*Videojuegos sociales para la asistencia y mejora de la pronunciación de la lengua española*)

- **Main researchers:** David Escudero-Mancebo and Valentín Cardeñoso-Payo
- **Start date:** 2015/01/01
- **End date:** 2018/06/30
- **Financing entity:** Ministerio de Economía y Competitividad (Spain)
- **Quantity:** 54,208€
- **Number of researchers:** 7
- **Project key:** TIN2014-59852-R
- **Summary.** TICs have become a key factor in the development and expansion of language learning during the last decade. This project is part of the research area related to CAPT systems and aims to study and develop a model of comparison between phrases that facilitates the development of innovative tools for the improvement of the pronunciation of Spanish as L2. It will be experienced some alternatives for automatic identification of improvable speech portions based on reference models that will also be developed as part of the results of this project. The learning methodology initially proposed is based on the user's performance of repetition of phrases or reading sentences presented by the system or generating short answers to questions or descriptions of scenes or situations, while receiving feedback that serve as a guide to improve the results. A social version enhances the motivation of the users. A solution based on a comparison and diagnosis module that incorporates the knowledge extracted from automatic comparison with reference speakers (native and non-native) is also proposed.

2. Title: Gamified software tools for pronunciation assessment and training (*Herramientas software ludificadas para la evaluación y entrenamiento de la pronunciación*)

- **Main researchers:** Valentín Cardeñoso-Payo
- **Start date:** 2018/01/01
- **End date:** 2020/12/31
- **Financing entity:** Consejería de Educación, Junta de Castilla y León (Spain)
- **Quantity:** 12,000€
- **Number of researchers:** 7
- **Project key:** VA050G18
- **Summary.** Recently, speech technology (an interdisciplinary subfield of computational linguistics traditionally dedicated to speech synthesis and recognition) is being applied to CAPT. The objective of this project is to deeply apply the knowledge of the research group about speech technology into the field of CAPT in two different domains: pronunciation training for people with intellectual disabilities, and L2 pronunciation training for learners. Since the research group has already developed projects along these lines, the funding requested in this project will serve to consolidate the research in these lines. The main results of the project to be carried out in the next three years will be: the design of an automatic prosody evaluation module for people with intellectual disabilities that serves to improve the Dashboard component of the video game already developed; the design of a module for evaluation and diagnosis of the quality of phoneme production and prosody in pronunciation training tools based on minimum pairs; We will focus mainly on Spanish and English as L2 languages. During the development of this project it is planned to complete two PhD theses and serve as support for applications for competitive research projects at national and European level.

3. Title: Automatic assessment of L2–Spanish pronunciation of native Japanese speakers (*Evaluación automática de la pronunciación del español como lengua extranjera para hablantes japoneses*)

- **Main researchers:** Valentín Cardeñoso-Payo
- **Start date:** 2014/01/01
- **End date:** 2017/12/31
- **Financing entity:** Consejería de Educación, Junta de Castilla y León (Spain)
- **Quantity:** 28,999€
- **Number of researchers:** 9
- **Project key:** VA145U14

- **Summary.** In this project it is proposed a solution that will combine techniques for generating and repeating speech fragments by non-native speakers with a comparison and diagnosis module that incorporates the knowledge extracted from automatic comparison with reference speakers. The target language is Spanish as a L2 and the preferred target population will be Japanese students. The collection of exercises will be elaborated based on the experience accumulated in the design of Spanish courses for Japanese speakers by the project researchers and in collaboration with other groups of research from three universities as with the *Asociación Europea de Profesores de Español*, and a company specialized in the development and commercialization of this type of systems. As a proof of concept, a prototype for mobile platforms will be developed to expand the locution database as it is used, facilitating the continuous improvement of the models.

4. Title: Predoctoral research fellowship of the University of Valladolid — 2015 (*Contratos predoctorales de la Universidad de Valladolid — Convocatoria 2015*)

- **Main researchers:** Cristian Tejedor-García
- **Start date:** 2016/09/01
- **End date:** 2020/08/31
- **Financing entity:** University of Valladolid (Spain)
- **Budgetary application:** 180113-541A.2.01-691
- **Quantity:** 70,000€
- **Summary.** The University of Valladolid is aware of the need to reinforce and complement the predoctoral phase in all areas of knowledge, facilitating that the best candidates can be trained in the different PhD programs offered by the University, and orienting their future toward research and teaching. This fellowship will allow the candidate to teach in the associated department a maximum of 60 hours per year. The main objective of this call is to increase the quantity and quality of new doctors, facilitating and promoting the realization of their PhD thesis within the framework of the best research projects. As a novelty, it is intended to encourage the candidates to obtain the PhD degree in three years by financing, in this case, a postdoctoral orientation period of an extra four year whose specific objective is the consolidation and improvement by the candidates of the knowledge acquired during the completion of their PhD theses. The conjunction of a candidate with good academic preparation who faces a quality doctoral training project within a quality research group, will undoubtedly ensure the future availability of doctors in the different academic disciplines that allow the University to have well-qualified candidates trained for a possible incorporation into their teaching and research staff.

5. Title: Predoctoral short-term fellowship of the University of Valladolid — 2019 (*Movilidad doctorandos ayudas para estancias breves en el desarrollo de tesis doctorales — Convocatoria 2019*)

- **Main researchers:** Cristian Tejedor-García
- **Start date:** 2019/08/26
- **End date:** 2019/11/26
- **Financing entity:** University of Valladolid (Spain)
- **Budgetary application:** 180113-541A.2.01-691
- **Quantity:** 2,746.67€
- **Host institution** Centre of Excellence in Educational Innovation (CEEI, University of Tallinn, Estonia).
- **Summary.** The main objective is to improve the analysis of the available game results of a CAPT learning game, COP, developed in one of the experimental prototypes of this thesis. This game establishes a competitive protocol to improve English pronunciation as a foreign language. COP has been designed to facilitate competitive mechanisms among players, launching and accepting challenges, and obtaining extra points or penalties depending on the results. A preliminary pronunciation improvement analysis has already been carried out, but much more information is available that can provide relevant data on the different player profiles, the game strategies adopted and the relationship of all this with the pronunciation improvement. In addition to explore new techniques of data analysis and time series to obtain results on player's behavior and their relationship with learning, it is also expected to obtain relevant information on those aspects of the game protocol that would need to be modified to improve future versions of the game.

Part V

Appendices

Appendix A

Minimal Pairs Lists Elaboration Algorithm

A.1 Algorithm Description

Elaborating a large list of minimal pairs for any language is not a trivial process and requires a considerable human effort. It is necessary to classify the words by their phonetic transcriptions and know their frequency of use in the language. However, it can be possible to reduce costs by using technology correctly. An algorithm for automatically elaborating minimal pairs lists from large amounts of text sources has been designed and tested. This algorithm is one of the steps of the protocol designed for including suitable minimal pairs for a CAPT system (see Figure 6.1 and Section 6.1.2) and it has been used and tested in the experimental prototypes of this dissertation, confirming to be valid for future experiments. It takes as **input** (1) a dictionary of words (orthographic–phonetic transcriptions) and (2) an unlimited quantity of text files. These files will determine the word frequency. The list of minimal pairs obtained in the **output** can be filtered and classified not only by the length and orthographic/phonetic transcriptions of the words, but also by their frequency of use in the text files provided.

A brute-force based algorithm in which all words (N) of length L are compared to each other could be valid with small amount of data. However, this approximation is not useful when dealing with large amount of data. In this thesis, a tree-based algorithm [211] which is more efficient than a brute-force based one has been designed. In particular, it takes into account not only N and L, but also the number of phonemes of the language, P. The time and space complexities of this algorithm are shown and compared to a brute-force based version in Table A.1.

	Brute-force based	**Tree-based**
Time	$\mathcal{O}(N^2 \cdot L)$	$\mathcal{O}(N \cdot L(P) \cdot L)$
Space	$\mathcal{O}(N \cdot L)$	$\mathcal{O}(P^{(L+1)} \cdot (L-1))$

TABLE A.1: Time and space complexities of the brute force-based and tree-based algorithm for elaborating minimal pairs lists. N refers to the number of words of a language with maximum length L of a source text and P is the number of phonemes of the language.

Algorithm 1 presents in pseudocode the main loop of the algorithm for elaborating the minimal pair lists. The methods that it requires are explained in Algorithms 2, 3, 4, 5, 6, and 7.

input : list of single words with their phonetic transcription (*dictFile*) and a set of texts from any data source of the same language (*frequencyFiles*).
output: list with the information of the minimal pairs displayed on the specified output.

```
/* See Section A.2 for details about INPUT and OUTPUT format   */
1  struct{
2      String orthographic;
3      String[] phonetic;
4      Integer frequency;
5  } WordInfo;
6  struct{
7      WordInfo word;
8      MinimalTree[] subTree;
9  } MinimalTree;
10 begin
       /* 1.  Preload of words                                  */
11     allWords[] ← WordFrequency(frequencyFiles);  /* See Algorithm 6 */
12     allWords[] ← PutPhonemes(dictFile, allWords); /* See Algorithm 7 */
13     a ← length of array allWords[];
       /* 2.  The algorithm starts                              */
14     L ← 45 ; /* Maximum length (number of phonemes) of the words */
15     P = [p1, p2, ..., pN] ;              /* Phonemes of the language */
16     base[] of MinimalTree ← InitBase(L);     /* See Algorithm 2 */
17     for w ← 0 to a do
18         word ← allWords[w];
19         o ← word.orthographic;
20         t ← word.phonetic[];
21         f ← word.frequency;
22         l ← length of array t[];
23         for r ← 0 to l do
24             tree ← SelectTree(base, l, r);          /* See Algorithm 3 */
25             leaf ← InsertWord(tree, o, t, f, r);    /* See Algorithm 4 */
26             if leaf ≠ null then
27                 PrintMinimalPairs(leaf, o, t, f, P);  /* See Algorithm 5 */
28             end
29         end
30     end
31 end
```

Algorithm 1: Main loop of the algorithm for elaborating minimal pairs lists.

```
1 Function InitBase(L) /* Creates an array which references to all
   possible trees.  There will be a set of trees for each possible
   word length.  That is, a tree for each possible rotation.    */
  Data:
      L: maximum length of the words in the search.
  Result: tree of the base array.
2 begin
3 |  return MinimalTree[size: L * (L + 1)/2]
4 end
```

Algorithm 2: InitBase() function of the algorithm for elaborating minimal pairs lists.

```
1 Function SelectTree(b, l, r) /* Finds the corresponding tree in the
   base array.                                                  */
  Data:
      b: base tree array,
      l: quantity of phonemes of the word,
      r: current rotation of the word.
  Result: array of trees
2 begin
3 |  return b[(l − 1) * l/2 + r]
4 end
```

Algorithm 3: SelectTree() function of the algorithm for elaborating minimal pairs lists.

```
1  Function InsertWord(t, str, phons, freq, r) /* Puts a word inside the
     specified tree if it is a new word.                                  */
   Data:
        t: tree of the base array,
        str: orthographic representation of the word,
        phons: array of phonemes of the word,
        freq: frequency of the word,
        r: current rotation of the word.
   Result: The leaf of the tree in which the word has been included (null in the
           case it has not been included).
2  begin
3      l ← length of array phons[];
4      for i ← 0 to l do
5          f ← phons[i];
6          if t.subTree[f] == null then
7              t.subTree[f] of MinimalTree ← MinimalTree[]
8          end
9          leaf ← t;
10         t ← t.subTree[f];
11     end
12     if t.word.orthographic == null then
13         t.word.orthographic ← str;
14         t.word.phonetic ← phons[];
15         t.word.frequency ← freq;
16         return leaf;                                  /* It is a new word */
17     end
18     else
19         return null;                              /* It is a repeated word */
20     end
21 end
```

Algorithm 4: InsertWord() function of the algorithm for elaborating minimal pairs lists.

```
1  Procedure PrintMinimalPairs(leaf, str, phons, freq, langPhons) /* Prints
     the information of the minimal pairs found.                    */
   Data:
        leaf: leaf of the tree to print the minimal pairs,
        str: word to look for its minimal pairs in the leaf,
        phons: array of phonemes of the word str,
        freq: frequency of the word str,
        langPhons: array of phonemes of the language.
   Result: The information of the minimal pairs found is printed to the
           specified output.
2  begin
3      NO_DIFF ← −1;
4      s ← length of string str;
5      total ← length of array langPhons;
6      for f ← 0 to total do
7          d ← NO_DIFF;
8          t ← leaf.subTree[f];
9          if t ≠ null then
10             phonemes ← t.word.phonetic[];
11             l ← length of array phonemes;
12             for (i ← 0 to l) and (d == NO_DIFF) do
13                 if phonemes[i] ≠ phons[i] then
14                     d ← i;
15                 end
16             end
17             if d ≠ NO_DIFF then
                   /* See Section A.2 for details about OUTPUT format
                      */
18                 print{output} : phonemes[d], phons[d], t.word.orthographic, str,
19                                 t.word.frequency+freq,t.word.frequency, freq, s,
20                                 phonemes, phons;
21             end
22         end
23     end
24 end
```

Algorithm 5: PrintMinimalPairs() procedure of the algorithm for elaborating minimal pairs lists.

```
1  Function WordFrequency(paths) /* Elaborates a dictionary of unique
     single words with their frequency of occurrence in the text
     files provided.                                              */
   Data:
       paths: list of unique names of the text source files in the file system.
   Result: list of unique single words with their orthographic transcription
           and frequency of apparition in the text files provided.
2  begin
3      SEP of Character ← SPACE ;                  /* Words divider */
4      list[] of WordInfo ← WordInfo[l];
5      total ← length of array paths[];
6      for f ← 0 to total do
7          file ← open(input(paths[f]));
8          fileData ← read(file);
9          close(file);
10         words[] ← fileData.split(SEP);
11         l ← length of array words[];
12         for i ← 0 to l do
13             str ← words[i];
14             if str ∉ list then
15                 word ← WordInfo;
16                 word.orthographic ← str;
17                 word.frequency ← 1;
18                 list.add(word);
19             end
20             else
21                 list[word].frequency ← list[word].frequency + 1;
22             end
23         end
24     end
25     return list[]
26 end
```

Algorithm 6: WordFrequency() function of the algorithm for elaborating minimal pairs lists.

```
1  Function PutPhonemes(path, words) /* Updates a list of single words
     with their phonetic transcriptions.                              */
   Data:
        path: unique name of the text source file in the file system. In each
              line it must appear the word separated by its phonetic
              transcription by a tab. The phonetic transcription's elements
              must be separated by spaces (see Section A.2 for more details),
        words: list of words to update their phonetic transcription.
   Result: list of words with their phonetic transcriptions updated.
2  begin
3      SEP of Character ← TAB ;          /* Transcriptions divider */
4      PHON_SEP of Character ← SPACE ;     /* Phonemes divider */
       /* 1.  Read all transcriptions from the source                  */
5      file ← open(input(path));
6      fileData[] ← readLines(file);
7      close(file);
8      l ← length of array fileData[];
9      dict[] of WordInfo ← WordInfo[size: l];
10     for i ← 0 to l do
11         blocks[] ← fileData[i].split(SEP);
12         word ← WordInfo;
13         word.orthographic ← blocks[0];
14         word.phonetic[] ← blocks[1].split(SEP_PHON)[];
15         if word.orthographic ∉ dict[] then
16             dict.add(word);
17         end
18     end
       /* 2.  Update the list of words with their transcriptions       */
19     l ← length of array words[];
20     for i ← 0 to l do
21         currentWord ← words[i];
22         str ← currentWord.orthographic;
23         if str in dict[] then
24             words[i].phonetic[] = dict[str].phonetic[];
25         end
26     end
27     return words[]
28 end
```

Algorithm 7: PutPhonemes() function of the algorithm for elaborating minimal pairs lists.

A.2 INPUT and OUTPUT Example

Given a list of words with their phonetic transcriptions and some text files, the algorithm calculates all possible combinations of word pairs that differ in a single phoneme indicating their frequency of apparition in the texts provided. For instance, a partial result of the execution of the algorithm described in Section A.1 for an English corpus training would be as follows:

INPUT1: *dictionary-EN.txt*

WORD	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**...**	**PN**
...	...	...	...	...	...	...	...	...
ABABA	AA1	B	AH0	B	AH0			
ABACHA	AE1	B	AH0	K	AH0			
ABACK	AH0	B	AE1	K				
ABACO	AE1	B	AH0	K	OW2			
ABACUS	AE1	B	AH0	K	AH0	S		
...	...	...	...	...	...	...	...	...

INPUT2: *book1-EN.txt*, *book2-EN.txt*, and *book3-EN.txt*

... TUB STANDING CHANNEL ACTIVITY SNAILS SHAKY CELERY ALOOF ASHAMED OBTAINABLE ANTS GRANDFATHER SPY MILK SPY NEED GRANDFATHER SPY MILK RECEIPT ABACK ABACHA HIDEOUS BOTTLE ABACUS STOMACH WORRY FRESH BLOT TITLE ABACK FLOAT BUTTON ABACK ANSWER RIGHT VERSED MAGIC PLATE MODERN WHIP TUB GRANDFATHER CHANNEL ACTIVITY CHANNEL SATELLITE JUSTICE GREGARIOUS CHICKEN DRIVER STUDENT BAN FABRICATE HUMAN BODY MAGIC CARRIER FERRY ADVOCATE VILLAGE NETWORK GRASS CHAUVINIST FRESH INSURANCE SWALLOW COALITION DUCK MUSICAL SUITCASE RESTRAIN MAGIC ECONOMIST MAGIC LOAN ELEGANT HOVER FILL LENGTH GRADUAL THEFT VIOLATION WEAVER MUSICAL SUITCASE RESTRAIN ...

OUTPUT: *minimalPairs-EN.txt*

D1	**D2**	**WORD1**	**WORD2**	**FT**	**F1**	**F2**	**L**	**W1P1**	**W1P2**	**W1P3**	**W1P4**	**W1PL**	**W2P1**	**W2P2**	**W2P3**	**W2P4**	**W2PL**
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
AE1	EH1	**ACTON**	**ECTON**	750	700	50	5	AE1	K	T	AH0	N	EH1	K	T	AH0	N
ER0	S	**ACTOR**	**ACTS**	1100	900	200	4	AE1	K	T	ER0	AE1	K	T	S		
K	F	**ACTOR**	**AFTER**	1850	900	950	4	AE1	K	T	ER0	AE1	F	T	ER0		
T	B	**ACTOR**	**AKBAR**	950	900	50	4	AE1	K	T	ER0	AE1	K	B	ER0		
K	N	**ACTOR**	**ANTAR**	1000	900	100	4	AE1	K	T	ER0	AE1	N	T	ER0		
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

INPUT1 entries refer to each line of the source file with the orthographic and phonetic transcription. The orthographic transcription appears in the first column (*WORD*); whereas in the rest of columns each phoneme of the phonetic transcription is represented (*P1 to PN*). A homogeneous appearance (text in uppercase or in lowercase) is also required.

The content of ***INPUT2*** is the raw text of the source files provided. A homogeneous appearance (text in uppercase or in lowercase) is also required.

Finally, each ***OUTPUT*** line refers to the information of the minimal pairs found. In particular, each phoneme of the minimal pair (the first two columns, *D1*, and *D2*) is also represented, the two words of the minimal pair (*WORD1* and *WORD2*), the number of times (frequency) each word appears in the source texts provided (columns *FT*, *F1*, and *F2*), the number of phonemes of both words (column *L*), and the list of phonemes for each word (from column *W1P1* to the end).

Appendix B

Experiments Comparative

In this Appendix several comparative tables related to the dimensions of the experimental procedure introduced in Chapter 6 are presented. The specific details of each experiment are described in Chapter 7. First, Table B.1 displays the effort and time needed to carry out the campaign of the experiments, recruit participants, and develop (planning, analysis and design, implementation, testing, and deployment) the prototypes [172]. Full-time development days of 7.5 hours are considered. This information could be used as a reference for conducting new experiments in the future or repeating them.

#Days	Minimal Pairs	TipTopTalk!	English Vowels	Japañol	COP
Development	30	90	75	45	55
Recruitment	7	6	1	1	6
Campaign	3	24	5	5	24

TABLE B.1: Number of development, recruitment, and testing days of each one of the prototypes of the experiments.

Second, Table B.2 compares the pronunciation training methodology followed in each experimental prototype (see Section 6.1 and Section 6.2 for more details).

	Activity selection	**Goal**	**Activities**	**L2 Minimal pairs**
Minimal Pairs	Guided	Training	Production *Right attempt: 1–5 positions*	20 en_US
TipTopTalk!	Free will	Training Competition	Exposure Perception Production *Right attempt: 1–5 positions (Variable)* Mixed	397 en_US 105 es_ES 168 cn_ZH 140 pt_BR 129 pt_PT 5 de_DE
English Vowels	Guided	Training	Theoretical Exposure Perception Production *Right attempt: 1st position* Mixed	72 en_US
Japañol	Guided	Training	Theoretical Exposure Perception Production *Right attempt: 1st position* Mixed	84 es_ES
COP	Free will Limited	Training Competition	Exposure Perception Production *Right attempt: 1–3 positions* Mixed	329 en_US

TABLE B.2: Comparative among experiments' training methodology.

Table B.3 compares the pronunciation assessment strategies carried out in each prototype of the experimentation (see Section 6.3 for more details).

	Perceptual tests	**Questionnaires**	**Focus group**	**Game assessment**
Minimal Pairs	-	Demographics	Random	Log files Production score
TipTopTalk!	-	Demographics UX	-	Log files Scores: game, production perception
English Vowels	Pre-test Post-test	Demographics UX	-	Log files Scores: game, production perception
Japañol	Pre-test Post-test	Demographics UX	-	Log files Scores: game, production perception
COP	-	Demographics UX Reasons for participating Intrinsic motivation Extrinsic motivation Pronunciation level self-concept Attitude toward competition Reasons for abandoning	Intrinsic motivation Extrinsic motivation Degree of competitiveness English Proficiency	Log files Scores: game, production perception

TABLE B.3: Comparative among experiments' pronunciation assessment approach.

Table B.4 shows the technology employed in each one of the prototypes of the experimentation (see Section 6.4 for more details).

	ASR	**TTS**	**Device**	**SO**	**Server**	**Analytics**
Minimal Pairs	Google-Android	Google-Android	Tablet Samsung SM-T800 (Speakers + Micro) (http)	MP app Android 6.0	Linux (pictures) Google (speech)	Local JSON
TipTopTalk!	Google-Android	Google-Android	Smart devices (Speakers + Micro) (http)	TTT app Android 2.3 or higher	Linux (pictures, logs and notifications) Google (speech, analytics, games)	Local JSON External JSON
English Vowels	Google-Android	Google-Android	PC (Speakers + Micro) (http)	EVow app Windows 7 NOX App 6	Linux (logs) Google (speech, analytics)	Local JSON External JSON
Japañol	Google Speech API (GCSTT) Kaldi	Google-Android	PC (Speakers + Micro) (http)	JÑ app Windows 7 NOX App 6	Linux (logs) Google (speechAPI, analytics)	Local JSON External JSON
COP	Google-Android Google Speech API (GCSTT)	Google-Android	Smart devices (Speakers + Micro) (http)	COP app Windows 7 NOX App 6	Linux (logs) Google (speechAPI, analytics)	Local JSON External JSON

TABLE B.4: Comparative among experiments' integrated technology.

Tables B.5 and B.6 present the gamified elements and strategies adopted for each prototype of the experimentation (see Section 6.6 for more details).

	Approach	**Leaderboards**	**Badges**	**Prize**
Minimal Pairs	Individualistic	No	No	Diploma Reward (focus group participants)
TipTopTalk!	Individualistic Competition (implicit)	Points and rounds: Total, en_US, cn_ZH, pt_BR, pt_PT, es_ES, de_DE	Trophies (points, languages, rounds) Motivational push messages	Diploma Reward (1st to 15th positions)
English Vowels	Individualistic	No	No	Diploma Reward
Japañol	Individualistic	No	No	Diploma Reward
COP	Individualistic Competition (explicit)	Points: Total	Trophies (points) Motivational push messages	Diploma Academic certification (60 or more completed challenges) Reward (1st to 15th positions)

TABLE B.5: Comparative among experiments' gamification instruments (I).

	Performance graph	**Avatar**	**Restrictions**	**Progress**
Minimal Pairs	No	No	Pronunciation: 420s all minimal pairs 5 attempts/word Difficulty level: easy	Final score
TipTopTalk!	Lesson's score Leaderboards	Profile picture	Discrimination: 10s/pair 1 attempt/pair Pronunciation: 60s 5 attempts/word Difficulty level: easy, normal, hard Clear tickets Right/Wrong perception word selection Adaptive round time	Final score Unlocking next lesson
English Vowels	Lesson's score	No	Discrimination: 10s/pair 1 attempt/pair Pronunciation: 60s/pair 5 attempts/word Difficulty level: hard	Final score Unlocking next lesson
Japañol	Lesson's score	No	Discrimination: 10s/pair 1 attempt/pair Pronunciation: 60s/pair 5 attempts/word Difficulty level: hard	Final score Unlocking next lesson
COP	Daily score Daily challenges Leaderboard	Profile picture	Discrimination: 10s/pair 1 attempt/pair Pronunciation: 60s/pair 3 attempts/word Difficulty level: normal 30 challenges maximum per day Similar matchmaking	Final score Challenge score Pending/ finished challenges score

TABLE B.6: Comparative among experiments' gamification instruments (II).

Table B.7 presents the main demographic characteristics of the participants in each one of the prototypes of the experimentation (see Section 6.7 for more details).

	Group features	**Age**	**Place and time**	**L1**	**L2**	**Origin**
Minimal Pairs	Natives English philology students Computer Engineering students	18-26	Lab. 7min/ 1 day	en_US es_ES	**53** en_US: *Natives*: 12 5W-7M *C1-C2*: 21 11W-10M *B1-B2*: 20 6W-14M	Spain USA
TipTopTalk!	Spanish and Chinese learners of English Spanish learners of Chinese *Others*	18-26 *No limit* *No limit*	Home 24 days	es_ES cn_ZH *Others*	**39** en_US (L1: es_ES): *C1-C2*: 15 8W-7M *B1-B2*: 15 8W-7M *A1-A2*: 9 5W-4M **4** en_US (L1: cn_ZH): *A1-A2*: 2 1W-1M **19** cn_ZH: *Natives*: 6 5W-1M *A1-A2*: 13 5W-8M *es_ES, pt_BR, pt_PT, de_DE*	Spain China *Rest of the world*
English Vowels	Spanish learners of English	18-26	Lab. 5 days	es_ES	**18** en_US: *B1-B2*: 18 10W-8M	Spain
Japañol	Japanese learners of Spanish	18-26	Lab. 5 days	jp_JP	**33** es_ES: *B1-B2*: 20 18W-2M *A1-A2*: 13 12W-1M	Japan
	Spanish professional speakers	25-50	Lab. 10 days	es_ES	**10** es_ES: *Natives*: 10 5W-5M	Spain
COP	Natives Spanish learners of English	18-60	Home 24 days	en_US es_ES	**354** en_US: *Natives*: 2 1W-1M *C1-C2*: 48 31W-17M *B1-B2*: 250 151W-99M *A1-A2*: 49 32W-17M *Staff*: 5 0W-5M	Spain USA

TABLE B.7: Comparative among experimentation participants' demographics. *W* and *M* refer to 'women' and 'men', respectively.

Table B.8 compares the corrective feedback (CF) strategies followed in each prototype of the experimentation (see Section 4.1 and Section 6.5 for specific details of each one of them).

	Implicit CF	**Explicit CF**
Minimal Pairs	Right/wrong answer sounds. Interface color change. Happy/sad smiley. Activity score.	Repetition's request of a mispronounced utterance with a recognized words message. Word synthesis.
TipTopTalk!	Word's phonetic transcription. Right/wrong answer sounds. Interface color change. Happy/sad smiley. Activity score.	Repetition's request of a mispronounced utterance with a recognized words message. Word synthesis. Dual listening to synthesized and own utterances.
English Vowels	Repetition's request of a mispronounced utterance without a recognized words message. Next exercise recommendation. Word's phonetic transcription. Right/wrong answer sounds. Interface color change. Happy/sad smiley. Activity score.	Theoretical-practical video. Word synthesis. Dual listening to synthesized and own utterances.
Japañol	Next exercise recommendation. Word's phonetic transcription. Right/wrong answer sounds. Interface color change. Happy/sad smiley. Activity score.	Theoretical-practical video. Repetition's request of a mispronounced utterance with a recognized words message and a pronunciation hint. Word synthesis. Dual listening to synthesized and own utterances.
COP	Word's phonetic transcription. Right/wrong answer sounds. Interface color change. Happy/sad smiley. Activity score.	Repetition's request of a mispronounced utterance with a recognized words message. Word synthesis. Dual listening to synthesized and own utterances.

TABLE B.8: Comparative among experiments' CF strategies.

Appendix C

Pre/Post-Tests of the Guided Learning Experiment

	Spanish	American English	
1	***la***	**la**	/lɑː/
2	***dan***	**Don**	/dɑːn/
3	***san***	**San**	/sæn/
4	***las***	**lass**	/læs/
5	***van***	**bun**	/bʌn/
6	***ven***	**Ben**	/ben/
7	***mes***	**mess**	/mes/
8	***das***	**Does**	/dʌz/
9	***sí***	**see**	/siː/
10	***sin***	**seen**	/siːn/
11	***fin***	**fin**	/fiːn/
12	***NIF***	**niff**	/nɪf/

	American English					
13	**Don**	/dɑːn/	**done**	/dʌn/		
14	**lock**	/lɑːk/	**lack**	/læk/		
15	**San**	/sæn/	**son**	/sʌn/		
16	**hat**	/hæt/	**hut**	/hʌt/	**hot**	/hɑːt/
17	**sack**	/sæk/	**suck**	/sʌk/	**sock**	/sɑːk/
18	**fen**	/fen/	**fan**	/fæn/		
19	**less**	/les/	**lass**	/læs/		
20	**men**	/men/	**man**	/mæn/		
21	**sit**	/sɪt/	**seat**	/siːt/		
22	**Tim**	/tɪm/	**team**	/tiːm/		
23	**San**	/sæn/	**son**	/sʌn/		
24	**teen**	/tiːn/	**tin**	/tɪn/	**ten**	/ten/
25	**peek**	/piːk/	**pick**	/pɪk/	**Peck**	/pek/

TABLE C.1: Pre-test and post-test words list of the English Vowels prototype.

	Spanish			
1	**caza**	/ˈkaθa/	**casa**	/ˈkasa/
2	**cocer**	/koˈθer/	**coser**	/koˈser/
3	**cenado**	/θeˈnaðo/	**senado**	/seˈnaðo/
4	**vez**	/beθ/	**ves**	/bes/
5	**zumo**	/ˈθumo/	**fumo**	/ˈfumo/
6	**moza**	/ˈmoθa/	**mofa**	/ˈmofa/
7	**cinta**	/ˈθiNta/	**finta**	/ˈfinta/
8	**concesión**	/koNθeˈsioN/	**confesión**	/koNfeˈsioN/
9	**fugo**	/ˈfuʏo/	**jugo**	/ˈxuʏo/
10	**fuego**	/ˈfweʏo/	**juego**	/ˈxweʏo/
11	**fugar**	/fuˈʏar/	**jugar**	/xuˈʏar/
12	**afuste**	/aˈfuʂte /	**ajuste**	/aˈxuʂte/
13	**pelo**	/ˈpelo/	**pero**	/ˈpɾo/
14	**hola**	/ˈola/	**hora**	/ˈoɾa/
15	**mal**	/mal/	**mar**	/maɾ/
16	**animal**	/aniˈmal/	**animar**	/aniˈmaɾ/
17	**hielo**	/ˈʤelo/	**hierro**	/ˈʤerro/
18	**leal**	/leˈal/	**real**	/rreˈal/
19	**loca**	/ˈloka/	**roca**	/ˈrroka/
20	**celada**	/θeˈlaða/	**cerrada**	/θeˈrraða/
21	**pero**	/ˈpɾo/	**perro**	/ˈperro/
22	**ahora**	/aˈoɾa/	**ahorra**	/aˈorra/
23	**enteró**	/ẽn̪teˈɾo/	**enterró**	/ẽn̪teˈrro/
24	**para**	/ˈpaɾa/	**parra**	/ˈparra/
25	**flotar**	/floˈtaɾ/	**frotar**	/fɾoˈtaɾ/
26	**flanco**	/ˈflaŋko/	**franco**	/ˈfɾaŋko/
27	**afletar**	/afleˈtaɾ/	**afretar**	/afɾeˈtaɾ/
28	**flotado**	/flotaˈðo/	**frotado**	/fɾotaˈðo/

TABLE C.2: Pre-test and post-test words list of the Japañol prototype.

The instructions given to the students in the pre-test and post-test are the following:

- Please read carefully the following list of word pairs (Table C.1 or Table C.2, respectively). Read them from top to bottom and from left to right.
- **You can read the word again if you think you have mispronounced it.**
- All words are accompanied by their phonetic transcription, in case you find it useful.
- You may read looking at the orthographic expression —*cat*— or at the transcription —/kæt/— but read the orthographic text at least one time.
- *(Only for Table C.1).* Notice that the first pairs consist of a Spanish word followed by an English word. The rest are English–English pairs. The list contains a few trios as well.

Appendix D

Speech Datasets

Table D.1 summarizes the main characteristics of the speech corpus gathered from the Japañol and COP prototypes. The non-native audio files of the Japañol prototype were derived from the pre/post tests and the production activities of the Exposure, Pronunciation, and Mixed modes of the CAPT tool. Native files were recorded under a controlled recording protocol. The audio files of the COP prototype were derived from the production activities of the Playing and Training modes of the training tool during nine days in three intervals of time (days 1–2–3, 11–12–13, and 22–23–24 of the competition).

	JAPANOL —Japañol prototype		*COP* —COP prototype	
Minimal pairs	84		329	
Unique words (total)	164 (168[1]) - pre/post: 55 (56[2])		591 (658[3])	
Word length (mean/MIN/MAX/SD)	4.29/2/8/1.07		5.61/2/11/1.70	
Unique phonemes	28		42	
Phoneme: frequency (%)	a: 16.9, o: 11.31, r: 9.0, e: 7.79, f: 5.27, s: 4.94, ɾ: 4.83, l: 4.5, t: 3.73, k: 3.73, u: 3.29, i: 3.29, θ: 3.29, n: 2.87, m: 2.41, ɣ: 1.87, j: 1.54, ð: 1.54, x: 1.32, b: 1.32, p: 1.1, d: 1.1, β: 0.88, w: 0.88, ŋ: 0.66, g: 0.33, ʤ: 0.22, z: 0.11		ɪ: 8.76, r: 6.93, t: 5.88, d: 5.43, s: 5.34, k: 5.15, l: 4.92, æ: 4.2, n: 4.2, b: 3.69, i: 3.28, p: 3.1, e: 3.01, ŋ: 2.87, ɛ: 2.51, ɑ: 2.37, ʌ: 2.37, w: 1.87, f: 1.69, a: 1.69, m: 1.69, g: 1.55, ʤ: 1.5, ʧ: 1.5, z: 1.46, h: 1.5, ə: 1.41, ɜ: 1.23, ʃ: 1.14, ʊ: 1.09, v: 1.0, u: 1.0, ɑː: 0.91, o: 0.87, ð: 0.64, j: 0.64, iː: 0.55, ɔ: 0.5, θ: 0.32, c: 0.09, y: 0.09, ʒ: 0.05	
	NATIVE	**NON-NATIVE**	**NATIVE**	**NON-NATIVE**
#Audio files (pre/post)	41,000 (0)	12,152 (2,800)	151 (0)	92,586 (0)
Time recorded (seconds/hours)	25,488.6/7.1	24,194/6.7	242/0.07	187,671.1/52.1
Size (megabytes)	5400	63	6.6	4,700
Audio quality (Khz/Kbps)	48/768	16/350	16/350	16/350
#File format	.wav	.flac .m4a	.flac .m4a	.flac .m4a
#Recording sessions (time/session)	1 (3h)	3 (1h), 2 (0.15h)	3 days (no limit)	9 days (no limit)
#Speakers)	10	33	2	354
#Female speakers	5	30	1	215
#Male speakers	5	3	1	137

[1] https://github.com/eca-simm/minimal-pairs-japanol-eses-jpjp
[2] See Table C.2
[3] https://github.com/eca-simm/minimal-pairs-cop

TABLE D.1: Descriptive statistics of the speech data gathered from the Japañol and COP prototypes. The symbol # means number of.

Appendix E

Kaldi ASR

E.1 Kaldi Directory Structure

Kaldi[1] is an open-source and free toolkit for speech recognition research [167]. It is based on Finite-State Transducers, FST (integrating the free software OpenFST[2]). The Kaldi project began in the workshop titled "Low Development Cost, High Quality Speech Recognition for New Languages and Domains" at Johns Hopkins University in 2009. The first version of the code was released on May 14th, 2011, as a subversion (svn)-based project at SourceForge. Relevant changes (four major versions and source code moved to GitHub) were applied until January 2017, when a stable version of the code was released (5.0.0) with a version number scheme (major/minor/patch). Kaldi is written in C++ and registered under the Apache License v2.0. Its code is supposed to be updated continuously by the community.

A detailed documentation is provided with scripts for building personalized *ad-hoc* recognition systems[3]. The Kaldi software distribution is available within a GitHub repository[4], and it is divided into five main folders[5]:

1. **egs**: example recipe scripts that allow to quickly build ASR systems for over 30 popular speech corpora, such as TIMIT, WSJ, TIDIGTS, or VOXFORGE (documentation is attached for each project). In this directory building your own systems is recommended.
2. **misc**: additional tools and supplies, such as HTK–to–Kaldi conversion scripts, and Kaldi scientific publications, not needed for proper Kaldi functionality.
3. **src**: Kaldi source code. It contains most of the source code for programs that the training recipes call.
4. **tools**: useful components and external tools, such as ATLAS and CLAPACK.
5. **windows**: tools for running Kaldi using Windows.

In this thesis the following structure of a Kaldi project has been carried out:

1. **run.sh**: main project file in which all commands are written. It is the starting point of the execution.

[1] https://github.com/kaldi-asr/kaldi
[2] http://www.openfst.org/
[3] http://kaldi-asr.org/doc/
[4] https://github.com/kaldi-asr/kaldi
[5] http://kaldi-asr.org/doc/kaldi_for_dummies.html

2. **cmd.sh**: file that specifies how and where (either online or locally) to run the CPU jobs.
3. **path.sh**: file which allows to run the project from any directory. It includes some features, such as Kaldi root directory, useful tools paths, the audio data directory, and specific variables of the project.
4. **conf**: folder that gathers files with configuration parameters for decoding and MFCC feature extraction processes.
5. **data**: contains relevant information data, such as transcripts, dictionaries, a lexicon, etc. for the *train* and *test* datasets. Audio files are usually placed into a sub-folder (*audio*).
6. **exp**: includes the output of the training and alignment scripts and the acoustic models.
7. **local**: contains other specific scripts for the project, such as the scoring method or the data split procedure.
8. **steps**: symbolic link to the *steps* directory in the base install. It includes essential scripts to train and decode ASR.
9. **utils**: symbolic link to the *utils* directory in the base install. It contains several scripts for data manipulation.

Figure E.1 summarizes the standard files and directories of a custom Kaldi project.

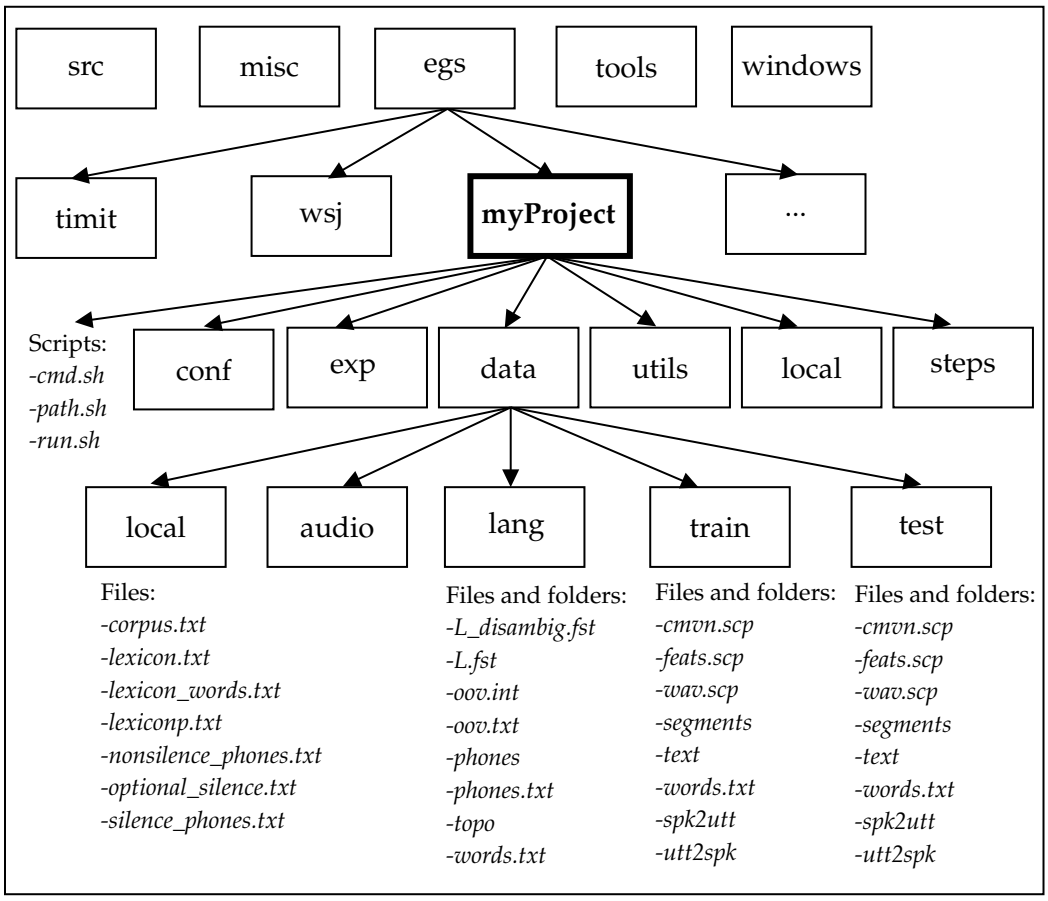

FIGURE E.1: Directory structure of a sample Kaldi project.

E.2 Elaborating an ASR System with Kaldi

Nine steps have been followed to elaborate, train, and obtain speech recognition results from a corpus with Kaldi (assuming a GMM/HMM framework) [167], [212]:

1. **Obtain speech data in audio files**. At least two datasets must be specified: *train* and *test*, and optionally a *validation* dataset. Audio files must be identifiable by speaker. The quantity of train files is a key factor to the final performance of the ASR system and to obtain reliable results. The length of the silences at the beginning and at the end of the audio files must be normalized.

2. **Prepare speech data transcriptions** (of the audio files) and optionally the start and end times of the sounds.

3. **Adapt and format speech data transcriptions**: list of phonemes and words of the transcripts, the language model, speaker identifications, and correspondences between audio files and transcriptions. The language model must be represented as an FST. Kaldi provides tools for converting language models in the standard ARPA[6] format to FSTs, such as SRILM[7].

4. **Extract acoustic features from the audio**. Standard MFCC and PLP features are obtained by commonly used **feature extraction** approaches, such as cepstral mean and variance normalization (CMVN), linear discriminant analysis (LDA) and semi-tied covariance (STC) / maximum likelihood linear transform (MLLT), among others (see step 8). Kaldi can be adapted not only to conventional **acoustic models**, such as diagonal GMMs with full covariance structures and subspace GMMs (SGMMs), but also to new kinds of model, such as DNNs. In Kaldi, context-dependent HMM states are represented by "pdf-ids".

5. **Train monophone models**. Monophone acoustic models are required as the first training step. These models do not include any contextual information about the preceding or following phone. They serve as a building block for triphone models giving contextual information.

6. **Align audio with the acoustic models**. Kaldi applies the Viterbi algorithm to optimize parameters of the acoustic model by cycling through training and alignment phases, realigning audio, and text. The alignment algorithms are speaker-independent.

7. **Train triphone models**. The triphone models represent a context-dependent trio of phonemes. Phonetic decision trees in Kaldi are efficient for arbitrary context sizes and also general enough to support a broad range of approaches. Kaldi decision-tree roots are more complex than the conventional approach in which each HMM-state of each monophone has a decision tree that asks questions about the left and right phones. Kaldi rots are shared among the phones and among the states of the phones, and questions can be asked about any phone in the context window, and about the HMM state.

8. **Re-align audio with the acoustic models & re-train triphone models**. In order to achieve better results, more training and alignment iterations are carried out. The main training and alignment algorithms supported by Kaldi are:

[6] https://cmusphinx.github.io/wiki/arpaformat/
[7] http://www.speech.sri.com/projects/srilm/

- **Delta + delta-delta**. These features are dynamic numerical estimates of the first and second order derivatives of the signal. Delta features are computed on the window of the original features. The delta-delta are then computed on the window of the delta-features.
- **LDA-MLLT**. The LDA algorithm builds HMM states with a reduced feature space from the feature vectors for all data. The MLLT training extracts a unique transformation for each speaker taking the LDA reduced feature space. It minimizes differences among speakers, being a step towards speaker normalization.
- **SAT-fMLLR/VTLN**. SAT, fMLLR, and VTLN stand for speaker adaptive training, feature space maximum likelihood linear regression and vocal tract length normalization, respectively. More homogeneous and standardized data is achieved with SAT by adapting to each specific speaker with a particular data transform. Instead of estimating variance using speaker or recording environment parameters, SAT applies its parameters to the phoneme. After SAT training, the acoustic model only contains speaker-normalized features. For alignment, the speaker identity is removed from the features by estimating the speaker identity. Both fMLLR and VTLN alignment algorithms can be applied.

9. **Evaluation of the results**. The last step is to analyze and evaluate the final decoding-graph. Kaldi decoders are native software classes that implement the core decoding algorithm. They do not need a particular type of acoustic model. The only require an object satisfying a very simple interface with a function that provides some kind of acoustic model score for a particular (input-symbol and frame). Therefore, Kaldi promotes the decoder's independence from the knowledge sources following a WFST framework [213]. They include the concept of "transition-id", that encodes a pdf-id, the phone which belongs to, and the transition (decoding graph arc) within the topology specification for that phone (an input label, an output label, and a weight/cost). Kaldi generates *lattices* with WFSTs. They represent the alternative word-sequences that are "sufficiently likely" for a particular utterance[8]. The final outcome of training and aligning with lattices generates a determinized and minimized HCLG decoding-graph transducer (combining four transducers to map from HMM states to word sequences): H stands for the HMM definitions. Its input symbols are WFSTs and its output ones represent context-dependent phones. C is the context-dependency. Its input symbols represent context-dependent phones and its output ones are phones. L represents the lexicon. Its input symbols are phones and its output ones are words. G is an *acceptor*. Both, its input and output symbols are the same. It encodes the grammar or language model. In order to give a score to the results, Kaldi scoring algorithm opens all the lattice files and obtain the best guess at the words in all of the utterances they contain with a confidence score. The WER is calculated with two files: the best estimation output in the previous step, and the correct labels. The program outputs the portion that matches with the target.

Steps 5, 6, and 7 are not necessary to be included in a **DNN training** procedure in Kaldi. It would require a previously trained GMM-HMM acoustic model and the phoneme-to-audio alignments (labeled frames) of the training audio. In fact, the

[8] http://kaldi-asr.org/doc/lattices.html

DNN is greatly affected by the quality of these previous results. The DNN acoustic model has some particular input and output nodes/layers, which correspond to the dimensions of the audio features and the labels that assign a class to those features, respectively. However, the number and size of the hidden layers are up to the researcher. As a conceptual basis, DNN training consists in comparing what the neural net predicted and what the real phoneme was. The results of assigning a phoneme label to frame with the phoneme labels from the previously trained GMM-HMM model. The net iterates over all of the training frames to adjust the weights and biases applying some loss function and backpropagation. DNN training does not require to train and align the transcriptions with the audio frames iteratively (in contrast to GMM-HMM). The detailed description of the three DNN training approaches with Kaldi exceeds the limits of this work. Interested readers can find an excellent explanation in depth at the official documentation web page[9] and in several publications, such as [214], [215].

[9] http://kaldi-asr.org/doc/dnn.html

Part VI

Bibliography

Bibliography

[1] Statista Research Department, *Number of network connected devices per person around the world from 2003 to 2020*, Nov. 2016, Statista. [Online]. Available: https://www.statista.com/statistics/678739/forecast-on-connected-devices-per-person. Accessed on: Apr. 14, 2019.

[2] B. E. Wiggins, "An Overview and Study on the Use of Games, Simulations, and Gamification in Higher Education", in *Gamification in Education: Breakthroughs in Research and Practice*, IGI Global, 2018, pp. 191–204. DOI: 10.4018/IJGBL.2016010102.

[3] D. W. Shaffer, K. R. Squire, R. Halverson, and J. P. Gee, "Video games and the future of learning", *Phi Delta Kappan*, vol. 87, no. 2, pp. 105–111, Oct. 2005. DOI: 10.1177/003172170508700205.

[4] C.-H. Chen, V. Law, and W.-Y. Chen, "The effects of peer competition-based science learning game on secondary students' performance, achievement goals, and perceived ability", *Interactive Learn. Environ.*, vol. 26, no. 2, pp. 235–244, 2018. DOI: 10.1080/10494820.2017.1300776.

[5] H. D. Brown and H. Lee, *Teaching by principles: An interactive approach to language pedagogy*. New York, NY, USA: Pearson Education, 2015, vol. 1. DOI: 10.2307/3587655.

[6] K. Beatty, *Teaching & Researching: Computer-Assisted Language Learning*. Routledge, 2013. DOI: 10.4324/9781315833774.

[7] Google Cloud, *Wavenet and other synthetic voices*, Dec. 2019, AI & Machine Learning Products. [Online]. Available: https://cloud.google.com/text-to-speech/docs/wavenet. Accessed on: Dec. 13, 2019.

[8] M. Meeker, *Internet trends 2017*, May 2017, Kleiner Perkins, Los Angeles, CA, USA, Rep. [Online]. Available: https://www.bondcap.com/report/it17. Accessed on: Mar. 14, 2018.

[9] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly, "A comparison of sequence-to-sequence models for speech recognition", in *Proc. Interspeech*, Stockholm, Sweden, Aug. 20–24, 2017, pp. 939–943. DOI: 10.21437/Interspeech.2017-233.

[10] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition", in *Proc. ICASSP*, New Orleans, NO, USA, Mar. 5–9, 2017, pp. 4845–4849. DOI: 10.1109/ICASSP.2017.7953077.

[11] 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. DOI: 10.1093/applin/amu076.

[12] W. Katz and P. Assmann, *The Routledge Handbook of Phonetics*, ser. Routledge Handbooks in Linguistics. Routledge, 2019, [Online]. Available: https://books.google.es/books?id=fa97swEACAAJ. Accessed on: Nov. 1, 2019, ISBN: 9781138648333.

[13] 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. DOI: 10.1017/S0958344008000724.

[14] H. Meng, "Developing Speech Recognition and Synthesis Technologies to Support Computer-Aided Pronunciation Training for Chinese Learners of English", in *Proc. 23rd Pacific Asia Conf. Lang. Info. Comput.*, City Univ. Hong Kong, China, Dec. 3–5, 2009.

[15] P. Avery and S. Ehrlich, *Teaching American English Pronunciation*. Oxford Univ. Press, 1995.

[16] D. T. Campbell and J. C. Stanley, *Experimental and quasi-experimental designs for research*. Ravenio Books, 2015.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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, Feedback Pronunciation Training Workshop*, ser. Feedback in Pronunciation Training Workshop, Saarland, Germany, Nov. 5–6, 2015.

[22] A. Rauber *et al.*, "TipTopTalk!: A game to improve the perception and production of L2 sounds", in *Abstr. New Sounds 8th Int. Conf. Second Language Speech*, Aarhus Univ., Aarhus, Denmark, Jun. 10–12, 2016, p. 160.

[23] 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. DOI: 10.1109/TLT.2020.2980261.

[24] 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. DOI: 10.21437/SLaTE.2017-5.

[25] 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. DOI: 10.21437/IberSPEECH.2018-21.

[26] 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.

[27] 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.

[28] 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.

[29] 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, [Online]. Available: https://cvc.cervantes.es/ensenanza/biblioteca_ele/aepe/pdf/revista_01_01_2018/revista_01_01_2018_33.pdf. Accessed on: Mar. 1, 2019.

[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. DOI: 10.12697/jeful.2019.10.1.05.

[31] 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. DOI: 10.1109/ACCESS.2020.2988406.

[32] R. L. Trask, *A dictionary of phonetics and phonology*. London, UK: Routledge, 2004.

[33] M. J. Munro and T. M. Derwing, "A prospectus for pronunciation research in the 21st century: A point of view", *J. Second Lang. Pronunciation*, vol. 1, no. 1, pp. 11–42, 2015.

[34] M. Levy and G. Stockwell, "CALL Dimensions: Options and Issues in Computer-Assisted Language Learning", *Modern Lang. J.*, vol. 91, no. 4, pp. 723–725, 2007. DOI: 10.1111/j.1540-4781.2007.00639_25.x.

[35] M. Rahimi, *Handbook of Research on Individual Differences in Computer-Assisted Language Learning*, 1st ed. IGI Global, 2015, ISBN: 9781466685208.

[36] C. Nagle, "Perception, production, and perception-production: Research findings and implications for language pedagogy", *World Lang. Cultures Publications*, no. 171, Aug. 2018.

[37] R. I. Thomson, "Computer-assisted pronunciation training: Targeting second language vowel perception improves pronunciation", *Calico J.*, vol. 28, no. 3, pp. 744–765, May 2011. DOI: 10.11139/cj.28.3.744-765.

[38] J. E. Flege, "Assessing constraints on second-language segmental production and perception", in *Phonetics and phonology in language comprehension and production: Differences and similarities*, A. Meyer and N. Schiller, Eds. 2003, pp. 319–355.

[39] E. Cámara-Arenas, *Native Cardinality: on teaching American English vowels to Spanish students*, ser. Historia y sociedad. Ediciones Univ. Valladolid, 2013, ISBN: 9788484487272.

[40] E. Cámara-Arenas, "The NCM and the Reprogramming of Latent Phonological Systems: A Bilingual Approach to the Teaching of English Sounds to Spanish Students", *Procedia - Social Behav. Sci.*, vol. 116, pp. 3044–3048, 2014.

[41] A. Baker, S. Goldstein, and P. Dolgin, *Pronunciation Pairs: An Introductory Course for Students of English. Student's Book*. Cambridge, UK: Cambridge Univ. Press, 1990.

[42] A. Baker, *Ship Or Sheep? Student's Book: An Intermediate Pronunciation Course*, 3rd ed. Ernst Klett Sprachen, 2006, vol. 1.

[43] J. D. O'Connor and C. Fletcher, *Sounds English-A pronunciation practice book*. Harlow, UK: Longman Group UK, 1999.

[44] A. Kukulska-Hulme, *Mobile-Assisted Language Learning*. Blackwell Publishing Ltd, 2012, ISBN: 9781405198431. DOI: 10.1002/9781405198431.wbeal0768.

[45] D. Escudero-Mancebo and M. Carranza, "Nuevas propuestas tecnológicas para la práctica y evaluación de la pronunciación del español como lengua extranjera", in *Proc. L Congreso Internacional de la Asociación Europea de Profesores de Español*, Burgos, Spain, Jul. 20-24, 2015, pp. 218–227.

[46] J. Levis, "Computer Technology In Teaching And Researching Pronunciation", *Annual Review App. Linguistics*, vol. 27, pp. 184–202, 2007. DOI: 10.1017/S0267190508070098.

[47] W. Li and D. Mollá-Aliod, "Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy", *Lecture Notes Comput. Sci.*, vol. 5459, 2009.

[48] M. Carranza, "Diseño de aplicaciones para la práctica de la pronunciación mediante dispositivos móviles y su incorporación en el aula de ELE", *El español entre dos mundos: Estudios de ELE en Lengua y Literatura*, pp. 279–297, 2014.

[49] J. Lee, J. Jang, and L. Plonsky, "The effectiveness of second language pronunciation instruction: A meta-analysis", *Appl. Linguistics*, vol. 36, no. 3, pp. 345–366, Jul. 2015. DOI: 10.1093/applin/amu040.

[50] 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. DOI: 10.1076/call.15.5.441.13473.

[51] B. S. Bloom, "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring", *Educ. Res.*, vol. 13, no. 6, pp. 4–16, 1984. DOI: 10.3102/0013189X013006004.

[52] D. Coniam, "Voice Recognition Software Accuracy with Second Language Speakers of English", *System*, vol. 27, no. 1, pp. 49–64, 1999.

[53] Tracey M. Derwing and Murray J. Munro and Michael Carbonaro, "Does popular speech recognition software work with ESL speech?", *TESOL Quarterly*, vol. 34, no. 3, pp. 592–603, 2000.

[54] F. Lys, "The development of advanced learner oral proficiency using iPads", *Lang. Learn. Technol.*, vol. 17, no. 3, pp. 94–116, Oct. 2013.

[55] B Pellom, "Rosetta Stone ReFLEX: Toward improving English conversational fluency in Asia", in *Proc. Int. Symp. Automatic Detection Errors Pronunciation Training*, Jun. 2012, pp. 6–8.

[56] M. El Tatawy, "Corrective feedback in second language acquisition", *Working papers TESOL Appl. Linguistics*, vol. 2, no. 2, pp. 1–19, Oct. 2002.

[57] A. Neri, C. Cucchiarini, and H. Strik, "Selecting segmental errors in non-native Dutch for optimal pronunciation training", *IRAL-Int. Rev. Appl. Linguistics Lang. Teaching*, vol. 44, no. 4, pp. 357–404, Dec. 2006. DOI: 10.1515/IRAL.2006.016.

[58] M. S. Mirzaei, K. Meshgi, and T. Kawahara, "Automatic Speech Recognition Errors as a Predictor of L2 Listening Difficulties", *CL4LC Workshop*, p. 192, Dec. 2016.

[59] M. Celce-Murcia and J. M. Goodwin, *Teaching Pronunciation*, 4th ed. London, UK: Thomson Learn., 2014, pp. 136–152.

[60] R. Akahane-Yamada, E. McDermott, T. Adachi, H. Kawahara, and J. S. Pruitt, "Computer-based second language production training by using spectrographic representation and HMM-based speech recognition scores", in *Proc. 5th ICSLP*, Sidney, Australia, Nov./Dec. 30–4, 1998, pp. 1–4.

[61] L. M. Tomokiyo, L. Wang, and M. Eskenazi, "An empirical study of the effectiveness of speech-recognition-based pronunciation training", in *Proc. 6TH ICSLP*, Beijing, China, Oct. 16–20, 2000, pp. 677–680.

[62] B. Mak *et al.*, "PLASER: Pronunciation learning via automatic speech recognition", in *Proc. HLT-NAACL Conf.*, Alberta, AB, Canada, May 27–Jun 1, 2003, pp. 23–29. DOI: 10.3115/1118894.1118898.

[63] R. Hincks, "Speech technologies for pronunciation feedback and evaluation", *ReCALL*, vol. 15, no. 1, pp. 3–20, Jun. 2003. DOI: 10.1017/S0958344003000211.

[64] R. Hincks, "Computer support for learners of spoken English", PhD thesis, Dept. Speech, Music Hearing, KTH Roy. Inst. Technol, Stockholm, Sweden, 2005.

[65] A. Neri, O. Mich, M. Gerosa, and D. Giuliani, "The effectiveness of computer assisted pronunciation training for foreign language learning by children", *Comput. Assisted Lang. Learn.*, vol. 21, no. 5, pp. 393–408, 2008. DOI: 10.1080/09588220802447651.

[66] D. Liakin, W. Cardoso, and N. Liakina, "Learning L2 pronunciation with a mobile speech recognizer: French /y/", *Calico J.*, vol. 32, no. 1, pp. 1–25, Jan. 2015. DOI: 10.1558/cj.v32i1.25962.

[67] 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. DOI: 10.21437/Interspeech.2018-34.

[68] G. Lord, "Podcasting communities and second language pronunciation", *Foreign Lang. Ann.*, vol. 41, no. 2, pp. 364–379, Mar. 2008. DOI: 10.1111/j.1944-9720.2008.tb03297.x.

[69] M. C. B. Alastuey, "Synchronous-voice computer-mediated communication: Effects on pronunciation", *Calico J.*, vol. 28, no. 1, pp. 1–20, Jan. 2010. DOI: 10.11139/cj.28.1.1-20.

[70] G. Smith, W. Cardoso, and C. G. Fuentes, "Evaluating text-to-speech synthesizers", in *Proc. 2015 EUROCALL Conf.*, Padova, Italy, Aug. 26–29, 2015. DOI: 10.14705/rpnet.2015.000318.

[71] H. Kataoka, M. Ito, and S. Yamane, "Retention of English sentences learned by reading aloud using text-to-speech (TTS) speech sounds: A longitudinal study in a Japanese high school", *International J. Res. Studies Educ. Technol.*, vol. 5, no. 1, Nov. 2015. DOI: 10.5861/ijrset.2015.1331.

[72] T. Bione, J. Grimshaw, and W. Cardoso, "An evaluation of text-to-speech synthesizers in the foreign language classroom: Learners' perceptions", in *CALL communities culture – short papers EUROCALL 2016*, Limassol, Cyprus, Aug. 2016, pp. 50–54. DOI: 10.14705/rpnet.2016.eurocall2016.537.

[73] Z. Handley, "Is text-to-speech synthesis ready for use in computer-assisted language learning?", *Speech Commun.*, vol. 51, no. 10, pp. 906–919, Nov. 2009. DOI: 10.1016/j.specom.2008.12.004.

[74] F. Soler Urzúa, "The acquisition of English /I/ by Spanish speakers via text-to-speech synthesizers: A quasi-experimental study", PhD thesis, Dept. Linguistics, Concordia Univ., Montreal, QC, Canada, 2011.

[75] C. Chapelle and J. Jamieson, *Tips for teaching with CALL: Practical approaches to computer-assisted language learning*. New York, NY, USA: Pearson Educ., 2008.

[76] C. A. Chappelle, "Innovative language learning: Achieving the vision", *ReCALL*, vol. 13, no. 1, pp. 3–14, 2001.

[77] D. F. Dalton, "Some techniques for teaching pronunciation", *Internet TESL J.*, vol. 3, no. 1, 1997.

[78] M. W. Tanner and M. M. Landon, "The effects of computer-assisted pronunciation readings on ESL learners' use of pausing, stress, intonation, and overall comprehensibility", *Lang. Learn. Technol.*, vol. 13, no. 3, pp. 51–65, Oct. 2009. DOI: 10125/44191.

[79] C. Fangzhi, "The teaching of pronunciation to Chinese students of English", *English Teaching Forum*, vol. 36, no. 1, pp. 37–39, 1998.

[80] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis", *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[81] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks", in *Proc. ICASSP*, Vancouver, BC, Canada, May 26–31, 2013, pp. 7962–7966.

[82] A. R. Bradlow, D. B. Pisoni, R. Akahane-Yamada, and Y. Tohkura, "Training Japanese listeners to identify English /r/ and /l/: Iv. Some effects of perceptual learning on speech production", *J. Acoustical Soc. America*, vol. 101, no. 4, pp. 2299–2310, Apr. 1997. DOI: 10.1121/1.418276.

[83] X. Wang, "Training Mandarin and Cantonese speakers to identify English vowel contrasts: Long-term retention and effects on production", PhD thesis, Dept. Linguistics, Simon Fraser Univ., Burnaby, BC, Canada, 2002.

[84] Y. Shinohara and P. Iverson, "High variability identification and discrimination training for Japanese speakers learning English /r/–/l", *J. Phonetics*, vol. 66, pp. 242–251, Jan. 2018. DOI: 10.1016/j.wocn.2017.11.002.

[85] R. I. Thomson, "Improving L2 listeners' perception of English vowels: A computer-mediated approach", *Lang. Learn.*, vol. 62, no. 4, pp. 1231–1258, Aug. 2012. DOI: 10.1111/j.1467-9922.2012.00724.x.

[86] N. C. Guilloteau, "Modification of phonetic categories in French as a second language: Experimental studies with conventional and computer-based intervention methods", PhD thesis, Dept. Psychol., Univ. of Texas Press, Austin, TX, USA, 1997.

[87] A. Weinberg and H. Knoerr, "Learning French pronunciation: Audiocassettes or multimedia?", *Calico J.*, vol. 20, no. 2, pp. 315–336, Jun. 2003. DOI: 10.1558/cj.v20i2.215-336.

[88] G. Lord, "(How) can we teach foreign language pronunciation? On the effects of a Spanish phonetics course", *Hispania*, vol. 88, pp. 557–567, Sep. 2005. DOI: 10.2307/20063159.

[89] P. Pearson, L. Pickering, and R. Da Silva, "The impact of computer-assisted pronunciation training on the improvement of Vietnamese learner production of English syllable margins", in *Proc. 2nd Pronunciation Second Lang. Learn. Teaching Conf.*, Iowa State Univ. Press, Ames, IA, USA, Oct. 7–8, 2011, pp. 169–80.

[90] E. M. Kissling, "Teaching Pronunciation: Is Explicit Phonetics Instruction Beneficial for FL Learners?", *Modern Lang. J.*, vol. 97, no. 3, pp. 720–744, 2013. DOI: 10.1111/j.1540-4781.2013.12029.x.

[91] D. M. Hardison, "Generalization of computer-assisted prosody training: Quantitative and qualitative findings", *Lang. Learn. Technol.*, vol. 8, no. 1, pp. 34–52, Jan. 2004.

[92] D. M. Hardison, "Contextualized computer-based L2 prosody training: Evaluating the effects of discourse context and video input", *Calico J.*, vol. 22, no. 2, pp. 175–190, Aug. 2005. DOI: 10.1558/cj.v22i2.175-190.

[93] Y. Hirata, "Computer-assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts", *Comput. Assisted Lang. Learn.*, vol. 17, no. 3–4, pp. 357–376, Aug. 2004. DOI: 10.1080/0958822042000319629.

[94] D. M. Chun, Y. Jiang, and N. Ávila, "Visualization of tone for learning Mandarin Chinese", in *Proc. 4th PSLLT Conf.*, Vancouver, British Columbia, Canada, Aug. 24-25, 2012, pp. 77–89.

[95] R. Hincks and J. Edlund, "Promoting increased pitch variation in oral presentations with transient visual feedback", *Lang. Learn. Technol.*, vol. 13, no. 3, pp. 32–50, Oct. 2009. DOI: 10125/44190.

[96] J.-Y. Lee, "The effects of pronunciation instruction using duration manipulation on the acquisition of English vowel sounds by pre-service Korean EFL teachers", PhD thesis, Dept. Linguistics, Univ. of Kansas, Lawrence, KS, USA, 2009.

[97] D. Liakin, W. Cardoso, and N. Liakina, "The pedagogical use of mobile speech synthesis (TTS): Focus on French liaison", *Comput. Assisted Lang. Learn.*, vol. 30, no. 3–4, pp. 325–342, Apr. 2017. DOI: 10.1080/09588221.2017.1312463.

[98] B. Luo, "Evaluating a computer-assisted pronunciation training (CAPT) technique for efficient classroom instruction", *Comput. Assisted Lang. Learn.*, vol. 29, no. 3, pp. 451–476, 2016. DOI: 10.1080/09588221.2014.963123.

[99] S. H. Yang and M. Chung, *Self-imitating feedback generation using gan for computer-assisted pronunciation training*, 2019. arXiv: 1904.09407 [cs.CL].

[100] D. Burleson, "Training segmental productions for second language intelligibility", PhD thesis, Dept. Linguistics, Indiana Univ. Press, Bloomington, IN, USA, 2007.

[101] G. Y. Eksi and S. Yesilcinar, "An investigation of the effectiveness of online text-to-speech tools in improving EFL teacher trainees' pronunciation", *English Lang. Teaching*, vol. 9, no. 2, pp. 205–214, Jan. 2016. DOI: 10.5539/elt.v9n2p205.

[102] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning", *Speech Commun.*, vol. 30, no. 2, pp. 95–108, Feb. 2000. DOI: 10.1016/S0167-6393(99)00044-8.

[103] C. Yarra, A. Srinivasan, S. Gottimukkala, and P. K. Ghosh, "SPIRE-fluent: A Self-Learning App for Tutoring Oral Fluency to Second Language English Learners", in *Proc. Interspeech*, Graz, Austria, Sep. 15–19, 2019, pp. 968–969.

[104] N. Moustroufas and V. Digalakis, "Automatic pronunciation evaluation of foreign speakers using unknown text", *Comput. Speech Lang.*, vol. 21, no. 1, pp. 219–230, Jan. 2007. DOI: 10.1016/j.csl.2006.04.001.

[105] V. Álvarez-Álvarez, D. Escudero-Mancebo, C. González-Ferreras, and V. Cardeñoso-Payo, "Evaluating Different Non-native Pronunciation Scoring Metrics with the Japanese Speakers of the SAMPLE Corpus", in *Proc. IberSPEECH*, Lisbon, Portugal, Nov. 23–25, 2016, pp. 205–214.

[106] M. Eskenazi and G. Pelton, "Pinpointing pronunciation errors in children's speech: Examining the role of the speech recognizer", in *ISCA Tutorial Resources Workshop Pronunciation Model. Lexicon Adaptation Spoken Language Technologies*, Colorado, CO, USA, Sep. 14–15, 2002, pp. 48–52.

[107] S. Pakhomov, J. Richardson, M. Finholt-Daniel, and G. Sales, "Forced-alignment and edit-distance scoring for vocabulary tutoring applications", in *Int. Conf. Text Speech Dialogue*, Brno, Czech Republic, Sep. 8–12, 2008, pp. 443–450. DOI: 10.1007/978-3-540-87391-4_57.

[108] N. Wiener, *Cybernetics: Control and communication in the animal and the machine*. New York, NY, USA: Wiley Online Library, 1948.

[109] R. L. Oxford, "Language learning styles and strategies", in *Teaching English as a Second or Foreign Language*, M. Celce-Murcia, Ed., Heinle & Heinle, 2001, pp. 359–366.

[110] H. Hattie John Timperley, "The power of feedback", *Rev. educational Res.*, vol. 77, no. 1, pp. 81–112, 2007.

[111] 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.

[112] R. Lyster and L. Ranta, "Corrective feedback and learner uptake: Negotiation of form in communicative classrooms", *Studies Second Lang. Acquisition*, vol. 19, no. 1, pp. 37–66, 1997. DOI: 10.1017/S0272263197001034.

[113] Y. Sheen and R. Ellis, "Corrective feedback in language teaching", *Handbook Res. Second Lang. Teaching Learn.*, vol. 2, pp. 593–610, 2011. DOI: 10.4324/9780203836507.ch36.

[114] K. N. Stevens, *Acoustic Phonetics*. The MIT Press, 2000, vol. 30.

[115] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, "Improving Non-native Mispronunciation Detection And Enriching Diagnostic Feedback With DNN-based Speech Attribute Modeling", in *Proc. ICASSP*, IEEE, Shanghai, China, Mar. 20-25, 2016, pp. 6135–6139.

[116] A. Jayakumar, M. Raghunath, M. Sakthipriya, S Akhila, A. Sadanandan, and P. Nedungadi, "Enhancing speech recognition in developing language learning systems for low cost Androids", in *Proc. ICCTICT*, IEEE, New Delhi, India, Mar. 11–13, 2016, pp. 80–84. DOI: 10.1109/ICCTICT.2016.7514556.

[117] C. Baur, J. Gerlach, M. Rayner, M. Russell, and H. Strik, "A shared task for spoken CALL?", *Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC)*, May 2016.

[118] H. Strik, J. Colpaert, J. van Doremalen, and C. Cucchiarini, "The DISCO ASR-based CALL system: Practicing L2 oral skills and beyond", in *Proc. 8th Int. Conf. Lang. Resour. Eval. (LREC)*, Istanbul, Turkey, May 21–27, 2012, pp. 2702–2707.

[119] A. Mitchell and C. Savill-Smith, *The use of computer and video games for learning: A review of the literature*. Learning and Skills Development Agency, 2004.

[120] Prensky, Marc, *Digital Game-Based Learning*. McGraw-Hill Pub. Co., 2004, p. 21, ISBN: 0071454004. DOI: 10.1145/950566.950596.

[121] M. D. Griffiths, "The educational benefits of videogames", *Educ. Health*, vol. 20, no. 3, pp. 47–51, 2002.

[122] K. D. Squire, "Video games and education: Designing learning systems for an interactive age", *Educ. Technol.*, vol. 47, no. 2, pp. 17–26, 2008.

[123] D. W. Johnson and R. T. Johnson, *Learning together and alone: Cooperative, competitive, and individualistic learning*, 5th ed. Boston: Allyn and Bacon, 1999, ISBN: 0205287719 9780205287710.

[124] S. Smith-Robbins, "This game sucks: How to improve the gamification of education", *EDUCAUSE review*, vol. 46, no. 1, pp. 58–59, 2011.

[125] É. Lavoué, "Towards Social Learning Games", in *Int. Conf. Web-Based Learn.*, Springer, 2012, pp. 170–179.

[126] P. J. Adachi and T. Willoughby, "Demolishing the Competition: The Longitudinal Link Between Competitive Video Games, Competitive Gambling, and Aggression", *J. Youth Adolescence*, vol. 42, no. 7, pp. 1090–1104, 2013. DOI: 10.1007/s10964-013-9952-2.

[127] D. R. Ewoldsen, C. A. Eno, B. M. Okdie, J. A. Velez, R. E. Guadagno, and J. DeCoster, "Effect of playing violent video games cooperatively or competitively on subsequent cooperative behavior", *CyberPsychol. Behav. Soc. Netw.*, vol. 15, no. 5, pp. 277–280, 2012. DOI: 10.1089/cyber.2011.0308.

[128] R. Van Eck and J. Dempsey, "The effect of competition and contextualized advisement on the transfer of mathematics skills a computer-based instructional simulation game", *Educ. Technol. Res. Develop.*, vol. 50, no. 3, pp. 23–41, Sep. 2002. DOI: 10.1007/BF02505023.

[129] N. E. Cagiltay, E. Ozcelik, and N. S. Ozcelik, "The effect of competition on learning in games", *Comput. Educ.*, vol. 87, pp. 35–41, Apr. 2015. DOI: 10.1016/j.compedu.2015.04.001.

[130] 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. DOI: 10.1177/0146167213520459.

[131] N. Storch, "Collaborative language learning", in *The Encyclopedia of Applied Linguistics*. American Cancer Soc., 2012, pp. 725–730, ISBN: 9781405198431. DOI: 10.1002/9781405198431.wbeal0153.

[132] R. McGloin, K. S. Hull, and J. L. Christensen, "The social implications of casual online gaming: Examining the effects of competitive setting and performance outcome on player perceptions", *Compt. Human Behav.*, vol. 59, pp. 173–181, 2016. DOI: 10.1016/j.chb.2016.02.022.

[133] P.-h. Tsai, "Computer-Assisted Pronunciation Learning in a Collaborative Context: A Case Study in Taiwan", *Turkish Online J. Educ. Technol.*, vol. 14, no. 4, pp. 1–13, 2015.

[134] A. F. AbuSeileek, "Cooperative vs. Individual Learning of Oral Skills in a CALL Environment", *Comput. Assisted Lang. Learn.*, vol. 20, no. 5, pp. 493–514, 2007. DOI: 10.1080/09588220701746054.

[135] S. Sampayo-Vargas, C. J. Cope, Z. He, and G. J. Byrne, "The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game", *Comput. Educ.*, vol. 69, pp. 452–462, Nov. 2013. DOI: 10.1016/j.compedu.2013.07.004.

[136] L.-J. Chang, J.-C. Yang, and F.-Y. Yu, "Development and evaluation of multiple competitive activities in a synchronous quiz game system", *Innov. Educ. Teaching Int.*, vol. 40, no. 1, pp. 16–26, 2003. DOI: 10.1080/1355800032000038840.

[137] P. J. Munoz-Merino, M. Fernandez Molina, M. Munoz-Organero, and C. Delgado Kloos, "Motivation and emotions in competition systems for education: An empirical study", *IEEE Trans. Educ.*, vol. 57, no. 3, Aug. 2014.

[138] S. Sepehr and M. Head, "Understanding the role of competition in video gameplay satisfaction", *Inf. Manage.*, vol. 55, no. 4, pp. 407–421, 2018. DOI: 10.1016/j.im.2017.09.007.

[139] M. Rayner, I. Strasly, N. Tsourakis, J. Gerlach, and P. Bouillon, "Menusigne: A serious game for learning sign language grammar", in *Proc. SLATE*, Stockholm, Sweden, Aug. 25–26, 2017, pp. 181–186. DOI: 10.21437/SLaTE.2017-32.

[140] J. ter Vrugte, T. de Jong, S. Vandercruysse, P. Wouters, H. van Oostendorp, and J. Elen, "How Competition and Heterogeneous Collaboration Interact in Prevocational Game-based Mathematics Education", *Comput. Educ.*, vol. 89, no. C, pp. 42–52, Nov. 2015. DOI: 10.1016/j.compedu.2015.08.010.

[141] W. Peng and J. Crouse, "Playing in parallel: The effects of multiplayer modes in active video game on motivation and physical exertion", *CyberPsychol. Behav. Soc. Netw.*, vol. 16, no. 6, pp. 423–427, 2013. DOI: 10.1089/cyber.2012.0384.

[142] H. Song, J. Kim, K. E. Tenzek, and K. M. Lee, "The effects of competition and competitiveness upon intrinsic motivation in exergames", *Compt. Human Behav.*, vol. 29, no. 4, pp. 1702–1708, 2013. DOI: 10.1016/j.chb.2013.01.042.

[143] C.-H. Chen, J.-H. Liu, and W.-C. Shou, "How competition in a game-based science learning environment influences students' learning achievement, flow experience, and learning behavioral patterns", *J. Educ. Technol. Society*, vol. 21, no. 2, pp. 164–176, 2018, [Online]. Available: http://www.jstor.org/stable/26388392. Accessed on: Mar. 17, 2018.

[144] S. Deterding, M. Sicart, L. Nacke, K. O'Hara, and D. Dixon, "Gamification. using game-design elements in non-gaming contexts", in *Proc. Conf. Human Factors Comput. Syst.*, Vancouver, BC, Canada, May 7–11, 2011, pp. 2425–2428, ISBN: 978-1-4503-0268-5. DOI: 10.1145/1979742.1979575.

[145] A. McFarlane, A. Sparrowhawk, Y. Heald, *et al.*, *Report on the educational use of games*. Cambridge, UK: TEEM, 2002.

[146] K. M. Kapp, *The Gamification of Learning and Instruction: Game-based Methods and Strategies for Training and Education*. John Wiley & Sons, 2012.

[147] D. Codish and G. Ravid, "Academic course gamification: The art of perceived playfulness", *Interdisciplinary J. E-Learn. Learn. Objects*, vol. 10, pp. 131–152, Jan. 2014. DOI: 10.28945/2066.

[148] D. Huynh and H. Iida, "An analysis of winning streak's effects in language course of "Duolingo"", *Asia-Pacific J. Inf. Technol. Multimedia*, vol. 6, no. 2, pp. 23–29, Dec. 2017.

[149] I. McGraw, B. Yoshimoto, and S. Seneff, "Speech-enabled card games for incidental vocabulary acquisition in a foreign language", *Speech Commun.*, vol. 51, no. 10, pp. 1006–1023, Oct. 2009. DOI: 10.1016/j.specom.2009.04.011.

[150] E. Danowska-Florczyk and P. Mostowski, "Gamification as a new direction in teaching Polish as a foreign language", in *Proc. 5th Int. Conf. ICT Lang. Learn.*, Florence, Italy, Nov. 15–16, 2012.

[151] D. Murad, R. Wang, D. Turnbull, and Y. Wang, "SLIONS: A karaoke application to enhance foreign language learning", in *Proc. 26th ACM Int. Conf. Multimedia*, Seoul, Republic of Korea, Oct. 13–16: ACM, 2018, pp. 1679–1687, ISBN: 978-1-4503-5665-7. DOI: 10.1145/3240508.3240691.

[152] P. Su, C. Wu, and L. Lee, "A recursive dialogue game for personalized computer-aided pronunciation training", *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 1, pp. 127–141, Jan. 2015. DOI: 10.1109/TASLP.2014.2375572.

[153] M. Sailer, J. U. Hense, S. K. Mayr, and H. Mandl, "How gamification motivates", *Comput. Hum. Behav.*, vol. 69, no. C, Apr. 2017.

[154] S. Schwab and J.-P. Goldman, "MIAPARLE: Online training for discrimination and production of stress contrasts", in *Proc. 9th Int. Conf. Speech Prosody*, Poznań, Poland, Jun. 13–16, 2018, pp. 572–576. DOI: 10.21437/SpeechProsody.2018-116.

[155] Z. Yuanyuan *et al.*, "Prosodic disambiguation by Chinese EFL learners in a cooperative game task", in *Proc. 9th Int. Conf. Speech Prosody*, Poznań, Poland, Jun. 13–16, 2018, pp. 979–983. DOI: 10.21437/SpeechProsody.2018-198.

[156] A. Berns, A. Gonzalez-Pardo, and D. Camacho, "Game-like language learning in 3-D virtual environments", *Comput. Educ.*, vol. 60, no. 1, pp. 210–220, 2013. DOI: 10.1016/j.compedu.2012.07.001.

[157] C. X. Wang, B. Calandra, S. T. Hibbard, and M. L. M. Lefaiver, "Learning effects of an experimental EFL program in Second Life", *Educ. Technol. Res. Develop.*, vol. 60, no. 5, pp. 943–961, 2012. DOI: 10.1007/s11423-012-9259-0.

[158] E. L. Deci and R. M. Ryan, "Overview of self-determination theory: An organismic dialectical perspective", *Handbook self-determination Res.*, pp. 3–33, 2002.

[159] E. D. Mekler, F. Brühlmann, A. N. Tuch, and K. Opwis, "Towards Understanding the Effects of Individual Gamification Elements on Intrinsic Motivation and Performance", *Comput. Hum. Behav.*, vol. 71, no. C, Jun. 2017.

[160] Y. Jia, Y. Liu, X. Yu, and S. Voida, "Designing leaderboards for gamification: Perceived differences based on user ranking, application domain, and personality traits", in *Proc. Conf. Human Factors Comput. Syst.*, New York, NY, USA, May 23–25: ACM, 2017, pp. 1949–1960, ISBN: 978-1-4503-4655-9. DOI: 10.1145/3025453.3025826.

[161] M. Celce-Murcia, D. M. Brinton, and J. M. Goodwin, *Teaching pronunciation: A course book and reference guide*, 2nd ed. Cambridge, UK: Cambridge Univ. Press, 2010.

[162] L. Daniele, *Estos son los idiomas extranjeros que más estudian los españoles*, Jan. 2017, ABC. [Online]. Available: https://www.abc.es/sociedad/abci-estos-idiomas-extranjeros-mas-estudian-espanoles-201701042125_noticia.html. Accessed on: Mar. 29, 2019.

[163] K. Saito and L. Plonsky, "Effects of Second Language Pronunciation Teaching Revisited: A Proposed Measurement Framework and Meta-Analysis", *Lang. Learn.*, vol. 69, no. 3, pp. 652–708, Apr. 2019. DOI: 10.1111/lang.12345.

[164] S. Loewen, "Focus on form", *Handbook Res. Second Lang. Teaching Learn.*, vol. 2, pp. 576–592, 2011.

[165] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001, ISBN: 0130226165.

[166] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O'Shaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors", *Speech Commun.*, vol. 55, no. 2, pp. 237–251, Feb. 2013. DOI: 10.1016/j.specom.2012.08.007.

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

[168] W. Yuxuan *et al.*, *Tacotron: Towards end-to-end speech synthesis*, 2017. arXiv: 1703.10135 [cs.CL].

[169] P. Taylor, *Text-to-speech Synthesis*. Cambridge, UK: Cambridge Univ. Press, 2009. DOI: 10.1017/CBO9780511816338.

[170] K. Hirose and J. Tao, *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*. Springer, 2015. DOI: 10.1007/978-3-662-45258-5.

[171] Google Cloud, *Cloud Text-to-Speech*, Dec. 2019, AI & Text-to-speech conversion powered by machine learning. [Online]. Available: https://cloud.google.com/text-to-speech/. Accessed on: Dec. 15, 2019.

[172] C. Larman and V. R. Basili, "Iterative and incremental development: A brief history", vol. 36, no. 6, 2003.

[173] L. M. Calvo-Magaz, *GETEPER: Aplicación multiplataforma para la gestión de tests perceptuales de audio*, Jul. 2019, Degree Final Project, Dept. Computer Science, University of Valladolid. [Online]. Available: http://uvadoc.uva.es/handle/10324/38760. Accessed on: Oct. 20, 2019.

[174] R. Sillero-Navajas, *Desarrollo de la modalidad multijugador para la aplicación Clash of Pronunciations*, Jul. 2017, Degree Final Project, Dept. Computer Science, University of Valladolid. [Online]. Available: http://uvadoc.uva.es/handle/10324/27645. Accessed on: Oct. 20, 2018.

[175] C. Tejedor-García, *TipTopTalk! Aplicación móvil para la mejora de pronunciación multilingüe mediante pares mínimos y gamificación*, Jul. 2016, Final Master's Degree Project, Dept. Computer Science, University of Valladolid. [Online]. Available: https://uvadoc.uva.es/handle/10324/17469. Accessed on: Oct. 20, 2018.

[176] A. Colina-Fernández, *Programación de aplicaciones Android para aprendizaje de idiomas*, Jul. 2018, Final Degree Project, Dept. Computer Science, University of Valladolid. [Online]. Available: http://uvadoc.uva.es/handle/10324/31297. Accessed on: Oct. 20, 2018.

[177] V. Y. Portero-López, *Videojuego serio para la promoción de Valladolid como ciudad para el aprendizaje del español como segunda lengua*, Jul. 2017, Final Degree Project, Dept. Computer Science, University of Valladolid. [Online]. Available: http://uvadoc.uva.es/handle/10324/25300. Accessed on: Oct. 20, 2018.

[178] J. Cal-Rioja, *Hound Word. Software para la mejora de la pronunciación en inglés*, Jul. 2016, Final Degree Project, Dept. Computer Science, University of Valladolid. [Online]. Available: http://uvadoc.uva.es/handle/10324/17963. Accessed on: Oct. 20, 2018.

[179] I. P. Association, I. P. A. Staff, *et al.*, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge, UK: Cambridge Univ. Press, 1999.

[180] T. J. Quirk, "One-Way Analysis of Variance (ANOVA)", in *Excel 2007 for Business Statistics: A Guide to Solving Practical Business Problems*, New York, NY, USA: Springer, 2012, pp. 165–181, ISBN: 978-1-4614-3734-5. DOI: 10.1007/978-1-4614-3734-5_8.

[181] A. Ross and V. L. Willson, "Paired samples t-test", in *Basic and advanced statistical tests*, Brill Sense, 2017, pp. 17–19. DOI: 10.1007/978-94-6351-086-8_4.

[182] M. L. McHugh, "The chi-square test of independence", *Biochemia Medica*, vol. 23, no. 2, pp. 143–149, Jan. 2013. DOI: 10.11613/BM.2013.018.

[183] 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. DOI: 10.1016/j.compedu.2012.04.015.

[184] P. E. McKnight and J. Najab, "Mann-Whitney U test", in *The Corsini Encyclopedia of Psychology*. American Cancer Society, Jan. 2010, ISBN: 9780470479216. DOI: 10.1002/9780470479216.corpsy0524.

[185] J. Cenoz and M. L. G. Lecumberri, "The acquisition of English pronunciation: Learners' views", *Int. J App. Linguistics*, vol. 9, no. 1, pp. 3–15, Apr. 2007. DOI: 10.1111/j.1473-4192.1999.tb00157.x.

[186] J. V. Casillas, "Production and perception of the /i/-/I/ vowel contrast: the case of L2-dominant early learners of English", *Phonetica*, vol. 72, no. 1, pp. 182–205, Apr. 2015. DOI: 10.1159/000431101.

[187] K. Nader, "Reconsolidation and the dynamic nature of memory", *Cold Spring Harbor Perspectives Biol.*, vol. 7, no. 10, pp. 1–20, Sep. 2015. DOI: 10.1101/cshperspect.a021782.

[188] 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.

[189] 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.

[190] M. Carranza, "Errores y dificultades específicas en la adquisición de la pronunciación del español LE por hablantes de japonés y propuestas de corrección", *Nuevos enfoques en la enseñanza del español en Japón*, C. Moreno and GIDE, Eds., Jan. 2012.

[191] S. Borsci, S. Federici, S. Bacci, M. Gnaldi, and F. Bartolucci, "Assessing user satisfaction in the era of user experience: Comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience", *Int. J. Hum. Comput. Interact.*, vol. 31, no. 8, pp. 484–495, 2015.

[192] R. E. Greenaway, "Chapter 13 — ABX Discrimination Task", in *Discrimination Testing in Sensory Science*, ser. Woodhead Publishing Series in Food Science, Technology and Nutrition, L. Rogers, Ed., Woodhead Publishing, 2017, pp. 267–288, ISBN: 978-0-08-101009-9. DOI: 10.1016/B978-0-08-101009-9.00013-7.

[193] H. Abdi, "The Kendall rank correlation coefficient", in *Encyclopedia of Measurement and Statistics*, Thousand Oaks, CA: SAGE Publications, Inc., 2007, pp. 508–510. DOI: 10.4135/9781412952644.n239.

[194] R. F. Woolson, "Wilcoxon signed-rank test", pp. 1–3, Sep. 2008. DOI: 10.1002/9780471462422.eoct979.

[195] J. De Mast, "Agreement and Kappa-type indices", *American Statistician*, vol. 61, no. 2, pp. 148–153, 2007. DOI: 10.1198/000313007X192392.

[196] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient", in *Noise Reduction in Speech Processing*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1–4, ISBN: 978-3-642-00296-0. DOI: 10.1007/978-3-642-00296-0_5.

[197] J. L. N. Alonso, J. M.-A. Lucas, and J. G. N. Izquierdo, "Validación de la versión española de la Échelle de Motivation en Éducation", *Psicothema*, vol. 17, no. 2, pp. 344–349, 2005.

[198] V. M. R. Vega, "Autoconcepto, ansiedad y motivación en el aula de idiomas", *MarcoELE: Revista Didáctica Español Lengua Extranjera*, no. 7, p. 8, 2008.

[199] J. Brooke *et al.*, "SUS—a quick and dirty usability scale", *Usability Eval. industry*, vol. 189, no. 194, pp. 4–7, Jun. 1996.

[200] P. E. McKight and J. Najab, "Kruskal-Wallis Test", in *The Corsini Encyclopedia of Psychology*. American Cancer Society, Jan. 2010, ISBN: 9780470479216. DOI: 10.1002/9780470479216.corpsy0491.

[201] Z.-H. Chen, "Exploring students' behaviors in a competition-driven educational game", *Compt. Human Behav.*, vol. 35, pp. 68–74, 2014. DOI: 10.1016/j.chb.2014.02.021.

[202] J. Koivisto and J. Hamari, "Demographic differences in perceived benefits from gamification", *Compt. Human Behav.*, vol. 35, pp. 179–188, Jun. 2014. DOI: 10.1016/j.chb.2014.03.007.

[203] H. N. Cheng, W. M. Wu, C. C. Liao, and T.-W. Chan, "Equal opportunity tactic: Redesigning and applying competition games in classrooms", *Comput. Educ.*, vol. 53, no. 3, pp. 866–876, Nov. 2009. DOI: 10.1016/j.compedu.2009.05.006.

[204] S. Deterding, "Eudaimonic design, or: Six invitations to rethink gamification", in *Rethinking gamification*, M. Fuchs, S. Fizek, P. Ruffino, N. Schrape, *et al.*, Eds., Meson Press, Jun. 2014, pp. 305–331. DOI: 10.25969/mediarep/727.

[205] B. Kollöffel and T. de Jong, "Can performance feedback during instruction boost knowledge acquisition? Contrasting criterion-based and social comparison feedback", *Interactive Learn. Environ.*, vol. 24, no. 7, pp. 1428–1438, 2016. DOI: 10.1080/10494820.2015.1016535.

[206] P. Bawa, "Retention in online courses: Exploring issues and solutions—a literature review", *SAGE Open*, vol. 6, no. 1, pp. 1–11, Jan. 2016. DOI: 10.1177/2158244015621777.

[207] T. W. Atchley, G. Wingenbach, and C. Akers, "Comparison course completion student performance through online traditional courses", *Int. Rev. Res. Open Distrib. Learn.*, vol. 14, no. 4, Sep. 2013. DOI: 10.19173/irrodl.v14i4.1461.

[208] M. Robinson, *Analytics Drive Design*, Feb. 2014, GamesAnalytics. [Online]. Available: https://www.gdcvault.com/play/1019527/Analytics-Driven. Accessed on: Dec. 29, 2018.

[209] M. Virvou and G. Katsionis, "On the usability and likeability of virtual reality games for education: The case of VR-ENGAGE", *Comput. Educ.*, vol. 50, no. 1, pp. 154–178, Jan. 2008. DOI: 10.1016/j.compedu.2006.04.004.

[210] Z. Zhao and D. Renard, "Viral promotional advergames: How intrinsic playfulness and the extrinsic value of prizes elicit behavioral responses", *J. Interactive Marketing*, vol. 41, pp. 94–103, Feb. 2018.

[211] D. Poojari, *Machine Learning Basics: Decision Tree From Scratch. Theoretical Framework*, Aug. 2019, Towards data science web page. [Online]. Available: https://towardsdatascience.com/machine-learning-basics-descision-tree-from-scratch-part-i-4251bfa1b45c. Accessed on: Sep. 28, 2019.

[212] E. Chodroff, *Corpus Phonetics Tutorial*, 2018. arXiv: 1811.05553 [cs.CL].

[213] M. Mohri, F. Pereira, and M. Riley, "Speech Recognition with Weighted Finite-State Transducers", in *Springer Handbook of Speech Processing*, Benesty, Jacob and Sondhi, M. Mohan and Huang, Yiteng Arden, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 559–584, ISBN: 978-3-540-49127-9. DOI: 10.1007/978-3-540-49127-9_28.

[214] D. Povey, X. Zhang, and S. Khudanpur, *Parallel training of DNNs with Natural Gradient and Parameter Averaging*, 2014. arXiv: 1410.7455 [cs.NE].

[215] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks", in *Proc. Interspeech*, Lyon, France, Aug. 25–29, 2013, pp. 2345–2349.