Universidad de Zaragoza

Departamento de Ingeniería Electrónica y Comunicaciones

Tesis doctoral

Personalización y Adaptación On-line a Trastornos y Variaciones de la Voz en Sistemas de Reconocimiento Automático del Habla

On-Line Personalization and Adaptation to Disorders and Variations of Speech on Automatic Speech Recognition Systems

Oscar Saz Torralba

Director de Tesis
Prof. Eduardo Lleida Solano

October 22, 2009

Go confidently in the direction of your dreams,
live the life you imagined.

Henry David Thoreau

Acknowledgements

Although it is my name the only one that appears as author in the front cover of this Ph.D. Thesis, I cannot forget of all the people that in different ways have participated through the long process of creation of this thesis. Here is a small (in comparison to all their influence) reminder for all of them.

First of all, I have to thank my family for their strong comprehension; especially my parents, Florencio and Adelaida for their patience during the long period that this thesis took until its completion.

Furthermore, this thesis is also the work of all the VIVOLAB group of the Grupo de Tecnologías de las Comunicaciones (Communications Technology Group) (GTC) at the Instituto de Investigación en Ingeniería de Aragón (Aragon Institute for Engineering Research) (I3A) in the University of Zaragoza. Of course, the most important influence is from my advisor, Prof. Eduardo Lleida, for all his good advices and his appreciative work. But this acknowledgment also extends to Prof. Enrique Masgrau, Dr. Alfonso Ortega, Dr. Luis Vicente, Dr. Luis Buera, Dr. Antonio Miguel, Juan Diego Rosas, Jesús Villalba, José Enrique García, David Becerril and Diego Castán. An important part of the work in this thesis was shared with W.-Ricardo Rodríguez and Carlos Vaquero, and I am really thankful to them for everything. From Carlos, I will never forget my period of tutoring his undergrad thesis, as well as the one of Antonio Escartín which have supposed a big influence in this work. Finally, Javier Simón from the Department of General and Hispanic Linguistics of the University of Zaragoza has been a extremely helpful colleague in the many subjects of this thesis.

The work in this thesis has been special to me because it has given me the chance to collaborate with different people and institutions dedicated to the education and assistance of handicapped people. Among them, special appreciation to the people from the Colegio Público de Educación Especial (Public School for Special Education) (CPEE) “Alborada”: José Manuel Marcos Rodrigo, César Canalís, Beatriz Martínez and Pedro Pegero. I have shared a lot of things with them all this time, and this thesis is also the result of their special interest in the new technologies applied to the handicapped. Other institutions that have collaborated in different ways with this work have been Coordinadora de Asociaciones de Personas con Discapacidad (Coordinating Committee of Associations of Handicapped People) (CADIS)-Huesca and Confederación Española de Federaciones y Asociaciones de Atención a las Personas con Parálisis y Afines (Spanish Confederation of Federations and Associations for the Assistance to People with Palsy and Related Handicaps) (ASPACE)-Huesca with Marta Peña, Verónica Bermúdez, Laura Abarca and Sara Mejuto as representatives of the whole team. I want to show my appreciation to all of them for their effort.

Special thanks also to Victoria Rodríguez for her very interesting work as teacher of Spanish for second language learners at the Vienna International School (VIS) and her interest in the results of this thesis.

I would like to make also a special reminder to the people of the Clinical Applications of Speech Technologies (CAST) group and the Speech and Hearing Group (SpandH) group from the Department of Computer Science at the University of Sheffield in Sheffield (United Kingdom), and to everybody in the Department of Electronic and Computer Engineering at McGill University in Montreal (Canada). They hosted me for the two summers of 2006 and 2007 respectively in what were really innovative experiences in all aspects. Prof. Phil Green and Dr. Stuart Cunningham from Sheffield helped me a lot in my first steps with their experience in the intensive work in speech technologies for the handicapped and introduced me to many new aspects in my research. Prof. Richard Rose from McGill provided me a very fruitful discussion for my work and I owe him a lot of my current knowledge in methods for confidence measuring and pronunciation evaluation.

Much of the work in this thesis was developed within the framework of the projects Dihana (TIC2002-04103-C03), Edecán (TIN2005-08660-C04) and SD-Team (TIN2008-06856-C05-04), funded by different agencies of the Spanish government. I also want to thank all the groups involved in these projects for their interest in this work. This includes the Pattern Recognition and Artificial Intelligence Group (GRFIA) at the Polytechnic University of Valencia lead by Dr. Emilio Sanchís and Dr. Encarnación Segarra, the Pattern Recognition and Speech Technologies Group (PR&ST) at the University of the Basque Country lead by Dr. María Inés Torres, the Speech Technologies Group (GTH) group at the Polytechnic University of Madrid lead by Dr. Rubén San Segundo and Dr. Javier Ferreirós and the people at the University of Alcalá de Henares lead by Dr. Javier Macías.

The final reminder is for all the people that I have met during all these years and that I can call my friends. This is a long list covering my old friends, former college mates and AEGEE people, all the people from the Department of Electronic Engineering and Communications of the University of Zaragoza, and my colleagues and friends from Sheffield and Montreal.

And finally, special appreciation to you, reader, because you are the only one who gives a meaning to this pages.

> Here lies the reader
> who will never open this book
> He is here forever dead
>
> - Milorad Pavić, *Dictionary of the Khazars*

Abstract

This thesis deals with the research and development of speech technology-based systems for the requirements of users with different impairments or disabilities, with the final aim of improving their quality of life. As these speakers usually present a wide range of speech disorders, their access to Automatic Speech Recognition (ASR)-based systems and similar is difficulted. The thesis proposes the use of personalization techniques to raise the performance of these speech-based systems in the proposed task of disordered speech.

This work performs all the steps in the research in speech technologies. A novel corpus containing nearly 3 hours of signal from young disabled speakers and nearly 9 hours of data from unimpaired age-matched individuals was acquired and it is to be described in the thesis. The collected data from unimpaired speakers is used for the development of a baseline ASR system adapted to young speakers. However, the baseline results achieved with this system by the impaired speakers are significantly degraded, compared to their unimpaired peers, pointing out the dramatic influence of the speakers' disorders in the performance of the ASR system.

From this starting point, a deep analysis of the disordered speech corpus is made in two directions. First one shows the acoustic degradation suffered by the speech uttered by the impaired speakers, compared to the control speech acquired from the unimpaired speakers. Later, speech and language disorders are proven to occur in the impaired speakers by means of analyzing the phonological and phonetic patterns in which the speakers are making their phoneme-level mispronunciations.

Proven that the disordered speech is degraded in both acoustic and lexical levels, acoustic and lexical adaptation to the speakers in the corpus are studied. Strong interrelations between both adaptation frameworks are observed and the need of matching both adaptation strategies is pointed out. Both adaptation frameworks (acoustic and lexical) make use of supervised data-driven techniques to provide the Word Error Rate (WER) improvement in the recognition, with a larger influence of the acoustic side in the ASR phase.

Given the impossibility to count at any time with labeled data when working with this kind of speakers, the need of developing a system that detect mispronunciations to avoid these acoustically inaccurate parts of the speech signal is required, prior to feed them to the adaptation systems. Traditional log-likelihood scoring and normalization trends in pronunciation verification are tested, altogether with some novel approaches. The possibility of identifying lexically correct and incorrect segments within the speech signal opens the gate for unsupervised adaptation frameworks.

These possibilities are studied over the different acoustic-lexical adaptation techniques used priorly. Finally, a proposal for on-line personalization is made, where the same utterances that the ASR has decoded are used for performing adaptation and create new models for the recognition of the following utterances from the speaker. In the end, the strong influence of the initial performance in ASR is observed, limiting the possibilities of application of these techniques.

The final part of the thesis covers all the attempts in the development of speech technology-based systems for the handicapped and speech therapy tools during this work. These systems are making use of the scientific knowledge acquired in this work and are open for all the community to use and share. The set of Computer-Aided Speech and Language Therapy (CASLT) tools in “Comunica”, result of a collaborative work, is shown to be successful in providing a semi-supervised aid for the speech handicapped with a great welcome by the community.

The scientific discussion and conclusions show that, even when there is still a great lack of knowledge in the use of speech technologies for disordered speech, there is an open possibility for the creation of personalized systems which can provide enhanced ASR performance to individuals with severe disabilities.

Resumen

Esta tesis se ocupa del la investigación y desarrollo de sistemas basados en tecnologías del habla para las necesidades de usuarios con discapacidades variadas, con el objetivo final de mejorar su calidad de vida. Dado que este tipo de usuarios generalmente sufren también algún tipo de trastorno en su habla y lenguaje, su acceso a sistemas de Reconocimiento Automático del Habla (RAH) u otros es más dificultoso. La tesis propone el uso de técnicas de personalización para mejorar los resultados en estos sistemas basados en el habla en la tarea que se propone en habla alterada.

La tesis lleva a cabo todos los pasos en la investigación en tecnologías del habla y reconocimiento del habla. Un nuevo corpus que contiene cerca de 3 horas de señal de jóvenes locutores con discapacidad y cerca de 9 horas de señal de locutores de la misma edad fué adquirido para este trabajo y es descrito en esta tesis. Los datos recogidos de locutores sin discapacidad son usados para el desarrollo de un sistema inicial de RAH adaptado al habla juvenil e infantil. Sin embargo, los resultados de base alcanzados por este sistema en los locutores con trastornos se ven significativamente degradados, comparados a los de los locutores sin discapacidad, señalando la importante influencia de los trastornos de los locutores en las prestaciones del sistema de reconocimiento.

Con este punto de partida, un análisis profundo del corpus de habla alterada se lleva a cabo en dos direcciones. La primera muestra la degradación acústica sufrida en el habla de los locutores discapacitados, comparada al habla de control de los locutores sin ningún tipo de discapacidad. Después, se prueba que el habla de los locutores discapacitados sufre variaciones léxicas importantes mediante el análisis los patrones fonológicos y fonéticos en los que los locutores realizan sus errores de pronunciación.

Una vez probado que este habla se ve degradada en los niveles acústico y léxico, se estudian situaciones de adaptación acústico-léxica a los locutores del corpus. La fuerte influencia entre ambos tipos de adaptación es una conclusión de estos trabajos y la necesidad de correlar ambas estrategias de adaptación se señala como necesaria para que realmente se obtenga la mejor mejora posible. Ambos técnicas de adaptación (acústica y léxica) hacen uso de algoritmos supervisados basados en datos para bajar la tasa de error de reconocimiento, donde la influencia de la parte acústica se ve como más relevante.

Dada la imposibilidad de contar en todos los casos con datos etiquetados cuando se trabaja con este tipo de locutores, se ve la necesidad de desarollar un sistema que detecte errores de pronunciación para evitar estos segmentos de voz acústica y léxicamente inexactos antes de usarlas como entrada a los sistemas de adaptación. Medidas de confianza basadas en scoring y técnicas de normalización de los mismos son estudiadas junto con aproximaciones más novedosas. La posibilidad de identificar las partes de la señal de voz fonéticamente correctas e incorrectas abre las puertas a la adaptación no supervisada. Estas posibilidades son exploradas sobre los diferentes sistemas de adaptación usados anteriormenta. Finalmente, se realiza una propuesta de perosnalización on-line, donde las propias seales que han sido decodificados por el sistema de reconocimiento se usan para realizar la adaptación a nuevos modelos que se usarán an las siguientes interacciones del usuario. Al final, se aprecia una gran influencia de la tasa de error inicial, limitando las posibilidades de aplicación de estas técnicas.

La parte final de la tesis cubre todos los esfuerzos realizados en el desarrollo de sistemas basados en tecnologías del habla para discapacidad y de herramientas logopédicas durante este tiempo. Estos sistemas hacen uso del conocimiento adquirido en este trabajo y están abiertos a toda la comunidad para su uso y distribución. Las herramientas para logopedia integradas en “Comunica”, como resultado de un largo trabajo de colaboración, muestran ser un caso de éxito en proporcionar una ayuda semi-supervisada para las personas con dificultades en el habla, habiendo tenido una gran acogida por parte de la comunidad.

La discusión científica y las conclusiones muestran que, incluso cuando aun queda mucho trabajo por hacer en este área, se debe poner un interés real en construir sistemas personalizados que se adapten a estas importantes variaciones de la voz, siendo posible el desarrollo de sistemas personalizados que provean de una interacción oral mejorada a individuos con discapacidad severa.

Contents

List of Figures

List of Tables

List of Acronyms

AAC Augmentative and Alternative Communication

APD Acoustic Phonetic Decoding

ARASAAC Portal Aragonés de la Comunicación Aumentativa y Alternativa (Aragonese Portal for the Augmentative and Alternative Communication)

ASPACE Confederación Española de Federaciones y Asociaciones de Atención a las Personas con Parálisis y Afines (Spanish Confederation of Federations and Associations for the Assistance to People with Palsy and Related Handicaps)

ASR Automatic Speech Recognition

AV@CAR Audio Visual at Car

BCI Brain-Computer Interface

CADIS Coordinadora de Asociaciones de Personas con Discapacidad (Coordinating Committee of Associations of Handicapped People)

CAI Caja de Ahorros de la Inmaculada

CALL Computer-Aided Language Learning

CAPT Computer-Aided Pronunciation Training

CARTV Corporación Aragonesa de Radio y Televisión (Aragonese Corporation of Radio and Television)

CASLT Computer-Aided Speech and Language Therapy

CATEDU Centro Aragonés de Tecnologías de la Educación (Aragonese Center for Technologies in Education)

CC Creative Commons

CEIP Colegio de Educación Infantil y Primaria (School for Primary and Infant Education)

CESEI Capítulo Español de la Sociedad de la Educación del IEEE (Spanish Area of the IEEE Education Society)

CPEE Colegio Público de Educación Especial (Public School for Special Education)

CPU Central Processing Unit

DPH Diputación Provicial de Huesca (Provincial Council of Huesca)

DVD Disordered Voice Database

EER Equal Error Rate

ELA Examen Logopédico de Articulación (Speech Therapy Exam for Articulation)

EM Expectation-Maximization

EU European Union

FAR False Acceptance Ratio

FFT Fast Fourier Transform

FP Framework Program

FR Fisher's Ratio

FRR False Rejection Ratio

GMM Gaussian Mixture Model

GOP Goodness Of Pronunciation

GRBAS Grade, Roughness, Breathiness, Aesthenia, Strain

GTC Grupo de Tecnologías de las Comunicaciones (Communications Technology Group)

HACRO Herramienta de Ayuda para la Confidencia de Realizaciones Orales (Help Tool for the Confidence of Oral Utterances)

HMM Hidden Markov Model

HTK Hidden Markov Model Toolkit

IASS Instituto Aragonés de Servicios Sociales (Aragonese Institute for Social Services)

ICSI International Computer Science Institute

IDCT Inverse Discrete Cosine Transform

I3A Instituto de Investigación en Ingeniería de Aragón (Aragon Institute for Engineering Research)

IES Instituto de Educación Secundaria (School for Secondary Education)

IPA International Phonetic Alphabet

ISLE Italian and German Spoken Learner's English

JFA Joint Factor Analysis

JNI Java Native Interface

KLD Kullback-Leibler Divergence

L1 First Language or mother tongue

L2 Second Language

LPC Linear Prediction Coefficients

LRE Language Recognition Evaluation

MAP Maximum A Posteriori

MFCC Mel Frequency Cepstral Coefficients

ML Maximum Likelihood

MLLR Maximum Likelihood Linear Regression

MMSE-LSA Minimum Mean-Square Error Log-Spectral Amplitude

NIST National Institute of Standards and Technology

NN Neural Network

OLP Ortho-Logo-Paedia

OLT Optical-Logo Therapy

OOV Out-Of-Vocabulary

OS Operating System

PDA Personal Digital Assistant

pdf probability density function

PER Phoneme Error Rate

PLON Prueba del Lenguaje Oral Navarra (Oral Language Test)

PV Pronunciation Verification

REAP Reader-Specific Lexical Practice for Improved Reading Comprehension

RFI Registro Fonológico Inducido (Induced Phonological Register)

RSD Relative Standard Deviation

SAMPA Speech Assessment Methods Phonetic Alphabet

SD Speaker Dependent

SGD Speech Generating Device

SI Speaker Independent

sKLD symmetric Kullback-Leibler Divergence

SLI Specific Language Impairment

SLM Statistical Language Model

SNR Signal-to-Noise Ratio

SPECO SPEech COrrector

SRE Speaker Recognition Evaluation

STARDUST Speech Training And Recognition for Dysarthic Users of aSistive Technology

SUSAS Speech Under Simulated and Actual Stress

Tball Technology-based assessment of language and literacy

TD Task Dependent

TI Task Independent

TIDigits Texas Instruments Digits

TIMIT Texas Instruments/Massachusetts Institute of Technology

T-norm Test Normalization

TTS Text-To Speech

UADatabase Universal Access Database

UBM Universal Background Model

UV Utterance Verification

VAD Voice Activity Detection

VHI Voice Handicap Index

VIS Vienna International School

VIVOCA Voice-Input Voice-Output Communication Aids

WER Word Error Rate

Z-norm Zero normalization

Chapter 1

Introduction

Longtemps, je me suis couché de bonne heure

- Marcel Proust, *À la recherche du temps perdu*

1.1 Introduction

The vision and treatment that society has given to handicapped and disabled individuals has widely changed as society itself has evolved during the ages. Recent anthropological discoveries in [Gracia et al., 2009] have shown up how Pleistocene tribal groups took care of heavily impaired individuals during childhood, attitude that was kept in earlier civilizations like the ancient Egypt in which some handicaps like blindness or achondroplasia [Kozma, 2005] were seen as realizations of the different deities and many of them were personally taken care of within the pharaoh's court. But Graeco-Roman civilization brought the dawn of the Western-culture prototype for beauty and perfection [Garland, 1995]. This prototype supposed the discrimination and rejection of all individuals that were outside this pattern of beauty. Hence, physical disabilities were seen as 'deformities' and these individuals did not have any rights as citizens and in some of the cases were condemned to exile or to a certain death.

Posteriorly, the rising of the main monotheist religions (Christianity and Islam) started changing concepts. The principles given in the holy books of both religions (Bible and Koran) appointed their believers to behave with 'charity' towards every handicap person [Berkson, 2005]. Social advances during the Renaissance era and the novel theories in physiology and other medical sciences started discovering the origins of the impairments and mentally handicapped individuals started having a legal coverage [Neugebauer, 1989]. Furthermore, the rationalism and social evolution with the industrial revolution made that families started taking care of their handicapped ones, although social inclusion was still not considered as a possibility. The view of the disabilities in the late Victorian era was pointed out with sharp precision in the description that Polish-born British author Joseph Conrad gave of Stevie, brother of the main character's wife in one of his master pieces: "The Secret Agent" [Conrad, 1907]:

For he was difficult to dispose of, that boy. He was delicate and, in a frail way, good-looking, too, except for the vacant droop of his lower lip. Under our excellent system of compulsory education he had learned to read and write, notwithstanding the unfavourable aspect of his lower lip. But as errand-boy he did not turn out a great success. He forgot his messages; he was easily diverted from the straight path of duty by the attractions of stray cats and dogs, which he followed down narrow alleys into unsavoury courts; by the comedies of the streets, which he contemplated open-mouthed, to the detriment of his employer's interests; or by dramas of fallen horses, whose pathos and violence induced him sometimes to shriek piercingly in a crowd, which disliked to be disturbed by sounds of distress in its quiet enjoyment of the national spectacle. When led away by a grave and protective policeman, it would often become apparent that poor Stevie had forgotten his address - at least for a time. A brusque question caused him to stutter to the point of suffocation. When startled by anything perplexing he used to squint horribly. However, he never had any fits (which was encouraging); and before the natural outbursts of impatience on the part of his father he could always, in his childhood's days, run for protection behind the short skirts of his sister Winnie.

The major change occurred at the end of World War II; the expansion of the so called 'welfare state' [Flora, 1987] to all the Western countries and the need to incorporate all the people who had suffered different injuries during the war lead to the development of modern mechanical aids like prosthesis and to the social inclusion of these individuals. The rapid increase in studies in several subjects like medicine, psychology and pedagogy produced an increase in knowledge of physical and psychical handicaps. The origin of several impairments got to be known and the new psycho-pedagogical theories pointed out that with correct attention, impaired individuals could play an important role in the society. Nowadays, the inclusion of the handicapped individuals is a major political issue and integration in the work market is a reality as long as everybody is aware of the capacities of each individual.

Finally, advances in information society in the last decades opened new gates for the handicapped world: Ubiquitous and immediate access to information without considering the personal conditions of each user. This seems the perfect scenario for disabled individuals to match the social capabilities of the unimpaired ones. However, this new gate has to be used carefully as it might create new barriers. For instance, the use of a computer still requires nowadays the correct handling of devices like the keyboard or a mouse. These devices rely on a fine control of the motor system, which is a new and heavy barrier for persons with major physical impairments.

1.2 Thesis Motivation

From this starting point that has been very briefly introduced, the motivation for this thesis arose due to the possibilities that speech technologies can have to proportionate new interface methods to interact with the new technological elements in our life or improve the communicative abilities of impaired individuals. However, these technologies are still not ready to provide an adequate performance in these domains. Systems based on Automatic Speech Recognition (ASR), for instance, are working finely these days in controlled situations like home control elements, hands-free devices for automotive environments or call centers. These environments share a controlled acoustic environment and a collaborative user with an acoustically normal, well-shaped speech. Unfortunately, this is not usually the case with users suffering from physical or developmental disabilities; as these impairments usually have associated physiological problems in the vocal tract or serious phonological and linguistic delays. A similar situation occurs in other cases with speech variants like dialectal speech or non-native speech, where the speaker is modifying the canonical form of the speech at the acoustic and lexical levels. Hence, while this thesis will be focusing on disordered speech, it expects to extract some type of generalizable results to similar tasks of speech variants.

Motivation for this work also came from the special interest that the staff members of the Colegio Público de Educación Especial (Public School for Special Education) (CPEE) “Alborada” have shown through the years for the development of technical aids for the handicapped [Martínez et al., 2007]; being fruitful in several devices like interactive communication boards [Negre, 2005][Negre et al., 2006], virtual mouse-control elements [Bergua, 2005][Bergua et al., 2006] or space-time orientation devices for the autistic [Falcó et al., 2006]. Their interest and offering for collaboration was the final cause that oriented this thesis to the work and objectives carried out finally.

As it is have been seen in this little presentation of motivations, and as it will be seen during the thesis, this work required of an effort for a multidisciplinary research. Knowledge in signal processing and speech technology was the basis of the thesis, as it was the main field it intended to cover. But a further knowledge in speech disorders and therapy, with special interest in special education, was necessary to understand the origins and consequences of these handicaps in the real world. Even psychology was a field of interest within the thesis, as these disorders are also related to the psychological processes of speech acquisition in children. Finally, another main component in this work was the linguistic and phonetic knowledge, as it was necessary to understand the already mentioned processes of speech production and acquisition.

1.3 Oral Communication Disorders

After the presentation of motivations for the thesis, this Section aims to provide a brief introduction in all the origins of oral communication disorders. Although the thesis deals mainly with disordered speech, it is important to place all possible types of impairments in the process of language acquisition and oral communication, because impaired speakers can show up other impairments in their speech and language. A small review on traditional speech and language therapy techniques will be furtherer provided.

1.3.1 Voice pathologies

Voice is the basic physiological part of oral communication. Speech production models [Fant, 1960] consider voice as a stimuli (glottal pulse) generated as an air flow in the lungs that becomes a periodic signal with the periodic movement (pitch frequency) of opening and closure of the vocal chords. This periodic stimuli is filtered by all the resonances of the elements and cavities of the vocal tract (tongue, palate, teeth, nose, ...) that shape the output speech signal as it is finally transmitted by the air. Unvoiced segments of speech like fricatives are generated by a burst of air created in the vocal tract.

Voice pathologies gather all kind of physiological impairments that prevent a correct creation of the glottal pulse in the lungs and trachea, a correct voicing creation in the vocal chords or a correct vocal tract configuration. These pathologies gather, among others:

- *Vocal chords lesions*, which are acquired voice pathologies in which an undesired element (like a polyp, cyst or nodule) appears on the vocal chords. This element limits the movements of opening and closure of the vocal chord during the speech production and, hence, the glottal pulse loses some of its properties, producing a loss in the quality of speech.

- *Dysphonia*, which produces abnormal changes in the voice, like sudden changes in pitch and volume or abnormal breathiness or raspiness. It can have different origins, ranging from bacterial elements like in laryngitis or by traumatic reasons like an intubation.

- *Morphological malformations*, like left clip and palate, produce a misshaping of one or some of the elements in the vocal tract. These morphological deviations produce the loss of articulatory properties of some sounds (for instance, nasalization of phonemes) due to the inability of the patient to correctly position the elements of the vocal tract.

- *Full larynx removal* due to cancer or other major traumatic situations can make the patient lose completely the phonation ability, unless rehabilitation teaches them to produce a glottal pulse directly with the esophagus (esophagical voice). However, the quality of the esophagical glottal pulse is poor and has a fundamental frequency abnormally low and with a very high variance which makes communication difficult.

Evaluation of the relevance of the voice impairment in the quality of life of the affected patients can be made with different subjective scales like the Voice Handicap Index (VHI) [Jacobson et al., 1997] or the Grade, Roughness, Breathiness, Aesthenia, Strain (GRBAS) scale [Hirano, 1981]. These scales rely on the ability of a practitioner to measure the patient's speech quality or in the subjective self-assessment of the patient.

1.3.2 Speech impairments

The difference between voice and speech is that speech carries a meaning that can be decoded and understood by a human listener. The smallest unit of speech is the phoneme, a sound that carries a certain information in a language (matched to a grapheme(s) in the written language). Syllables and words are the next step in the generation of speech, being the latests the smallest units with a full meaning. People with speech disorders suffer, hence, from an inability to articulate properly certain sounds and phonemes, or a sequence of them, resulting in a difficulty to utter full words correctly and understandably.

Distinction of different speech impairments is hard to set, as there might be different elements appearing at the same time to produce a given effect in the patient's speech. In general, a speech impairment producing a mispronunciation of one or several phonemes is gathered under the term of dyslalia [Pascual-García, 1992]. Four types of mispronunciations are distinguished: *Substitutions*, in which another phoneme, simpler to pronounce for the speaker is pronounced in the place of the correct phoneme. *Deletions*, in which the correct phoneme is erased from the pronunciation of the patient. *Insertions*, in which an extra phoneme appears intercalated in the pronunciation. *Distortions*, in which the phoneme is incorrectly pronounced, but it does not resemble another phoneme of the patient's language.

A full distinction of dyslalias [Perelló, 1984] can be made attending to the origin of the disorder:

- *Evolutive dyslalia*, which appears when a child old enough to produce correct speech (above 4-5 years) is still producing the patterns of mispronunciations of children in earlier ages [Bosch-Galcerán, 2004]. This dyslalia is not related to any neurological or morphological problem and is usually caused by a phonological delay in the children due to learning problems. Speech therapy is required, hence, to help the patient to achieve an speech ability matched to the child's age.

- *Functional dyslalia*, which is caused by an abnormal function of the organs involved in the articulation of the sounds but without existing a morphological impairment on those organs. The origin usually is an immaturity of the patient for the articulation process due to lack of motorize ability, hearing loss or psychological or environmental factors. Speech therapy is also of major help in moderate cases to improve the communicative skills of the patient. This disorder is linked to cognitive and development disorders which limit the neurological and learning capabilities of the patient.

- *Audiogenous dyslalia*, which is caused indirectly by a hearing loss. Lack of correct perception and feedback makes the patient unable to discriminate different sounds and phonemes; as most of the speech acquisition is based on the ability of the child to repeat the sounds heard in the environment. This inability will produce a further delay in the correct acquisition of these phonemes. Speech therapy is required in these cases, but it has to be oriented to the special needs of these patients, as auditive reinforcement shall not work for them.

- *Organic dyslalia*, which is caused by alterations in any of the physiological elements involved in the articulation of phonemes. Two of the affections related to organic dyslalias which are more common are dysarthria and dysglossia: *Dysarthria* is originated by a major impairment in the nervous system. It is one of the most severe speech affections and it is usually related to brain damage like cerebral palsy or stroke. Rehabilitation via speech therapy is usually of little help as long as the neurological disorder is still present in the patient. *Dysglossia* is originated by serious malformations in the organs of the vocal tract involved in articulation (tongue, palate, lips, teeth, ...). These malformations (genetic or acquired) produce an inability to generate the sounds in which the affected organ has a major role. Left clip and palate is a typical disorder producing dysglossia, serving also as example of voice pathology producing an speech disorders. When correction of the organic impairment is possible (usually via surgery), speech therapy is required to help the patient recover a proper speech.

1.3.3 Language impairments

Language is all the process in the transmission of ideas from one human being to another. It gathers all the elements presented previously like voice and speech, but when referring to language impairments, they are specifically all oral communication disorders related to the psychological process of language (symbolization of language, formulation and creation of ideas, etc) [Launa and Borel-Maisonny, 1989].

Language disorders are usually gathered under the Specific Language Impairment (SLI) framework [Aguado, 1999]. Language impairments are difficult to detect and diagnose because they are not connected to any effect in the physiological or psycho-sociological causes of other impairments. At this point, it is important to distinguish between language delays and language impairments. A language delay usually supposes that the child has an expressive language corresponding to children of younger ages and can be recover relatively easily with therapy.

However, language impairments, in the forms of dysphasia (incorrect language) or aphasia (lack of language), affect all three elements of language [Rapin and Allen, 1983]:

- *Expressive language* is affected when the articulation does not match the correct production of sounds in the language of the patient.

- *Comprehensive language* fails when there is a difficulty in comprehending oral language by the patient, without any difficulty in the production of the own language.

- *Central processes of language* are affected when there is an inability to produce language because the patient is unable to evocate the ideas to be expressed orally; oral production can be good, but only with memorized sentences or sentences without an understandable meaning.

1.3.4 Language acquisition and therapy

The natural process of language acquisition starts in the first year of the infant's life when the child starts uttering the first vocalic sounds to create the first syllables and words [Bosch-Galcerán, 2004]. In this period, it is very important that the child can perceive normally the environment because listening is the way in which the acquisition of the first sounds will be made.

After the first year, the first complete meaningful words appear in the child's speech. They are based in simple vocalic structures like Consonant-Vowel (CV), where the consonants are mostly plosives and nasals. The rest of phonemes (fricatives, vibrants or liquids) are more complicated due to the complex articulation movements that they suppose and are not fully achieved till the 4th-5th year of life (when the average rate of mispronunciations gets below 20%). Another complex structures like consonant clusters or dypthongs are also not mastered till the 4th year of life when they start being correctly pronounced in a consistent way. In parallel, the child learns the syntax and semantics to create meaningful sentences, while increasing the vocabulary [Moore and ten Bosch, 2009].

Phonological evaluation and phonological registers, where a child utters a set of specific words in front of a professional speech therapist, are the way in which the phonological advances of a student are assessed; and when the development of the speech is below the levels of quality of the age-matched children, it becomes evident the need of professional help and speech therapy.

In these cases or when a delay in language acquisition is detected, speech therapy is required to control and overcome those speech and language difficulties. Providing speech therapy is a hard and time demanding task which, unfortunately, does not always count with the sufficient resources to reach all the children that require it in the schools. As it has been seen previously, there are very different disorders that can affect a child's speech, limiting the student's performance in school or in the familiar and social environments. Furthermore, speech therapy becomes a difficult task when dealing with so many possible disorders [Bustos, 1995].

When children of very young age show an early difficulty in the first stages of their oral production, prelinguistic therapy is provided according to multiple handbooks [Acero-Villán and Gomis-Cañete, 2005]. The activities proposed in these cases aim to provide the young patient with reinforcement and feedback on the possibilities of speech production. Training of breathing, voicing and vocalization are done in personalized speech therapy sessions with a professional therapist who stimulates the oral production of the patient.

The phonological level of the language is traditionally treated by evaluations in which the patient is asked to utter a set of words in which certain phonemes appear in certain positions and context. The most popular of these handbooks for Spanish are the Examen Logopédico de Articulación (Speech Therapy Exam for Articulation) (ELA) [Albor, 1991], the Prueba del Lenguaje Oral Navarra (Oral Language Test) (PLON) [Aguinaga et al., 2004], the Registro Fonológico Inducido (Induced Phonological Register) (RFI) [Monfort and Juárez-Sánchez, 1989] as well as other proposals [Bosch-Galcerán, 2004]. Posteriorly, the therapy consists of personal sessions with the therapist where the patient is encouraged to utter a set of words selected according to individual needs and reinforced on the correct pronunciation and the correct positioning of the vocal tract elements.

Concerning higher levels of the language, more related to the communicative or functional side of language, therapy consists on personal interaction between patient and therapist. The therapist presents a certain situation of the real life and asks for the oral solution that the patient would give to it. Several handbooks are published by different authors to provide these practical situations [Monfort and Monfort-Juárez, 2001a, Monfort and Monfort-Juárez, 2001b] to the community.

1.4 Objectives and Methodology

This section brings together the goals and objectives set for this thesis. Main objectives are in the scientific field and related to the widening in the knowledge in speech technologies (mainly ASR and speech assessment) for variants of speech (specifically speech disorders) with the final goal of using this knowledge for the present and future development of speech-based technical aids for the handicapped. The development of possible tools for this purpose and for speech therapy based on robust speech assessment is also seen as a separate set of objectives. Finally, the methodological procedure followed in the thesis will be presented.

1.4.1 Scientific objectives

There are three main scientific objectives related to the present thesis:

The first objective of the thesis is to acquire a *fully functional speech corpus* containing speech from different speakers suffering from a wide range of speech and language disorders. The speech signals and the transcriptions of the uttered words are the basis and first requirement for the completion of this thesis. A deep analysis of the corpus has to be made in order to be able to understand the properties of this kind of speech prior to start the researching in techniques to improve the ASR performance in this task.

A further objective is to study algorithms for *adaptation techniques for ASR* in the presence of speech variants and for the detection and assessment of these situations. Speaker adaptation has been shown to provide robust performance improves in ASR systems in situations of normal healthy speech; but studies of new challenges to be considered when dealing with disordered speech have to be part of this thesis and proposals to overcome these possible gaps will have to be made.

The third and last objective is the research in algorithms for *assessment of the speech proficiency* of speech impaired individuals. Assessment and detection of pronunciation mistakes is an important issue in the work with disorders of speech as it helps for diagnosis and rehabilitation of the speaker and, furtherer, it can provide valuable information to other speech-based systems like ASR to improve their performance or provide unsupervised frameworks of use.

1.4.2 Development objectives

The possible development of real-world applications that can help handicapped persons in different aspects of their quality of life is fixed as another set of objectives for this thesis. This goal will require of collaboration and support from educative or assistance centers to the handicapped, as their knowledge on the real needs of the impaired population and their experience in the work with them is strongly necessary to develop fully effective tools. Two types of elements will be considered:

First, *devices for interaction and control* of the environment partly or fully based on voice or speech. These devices aim to take an oral input from the user (combined with other possible inputs) and use it for the control of elements like computers or household systems. The oral input will be analyzed either for speech recognition or keyword spotting systems or for different voice features that can discriminate different actions to take.

Moreover, tools for *Computer-Aided Language Learning (CALL)* benefit from robust speech assessment techniques like the ones that are aimed to be studied in this thesis. The development of CALL tools is, hence, another objective for this thesis. These tools should take the speech input from the user, evaluate it according to the likeliness to the prompted word or sentence and provide the user with feedback on this evaluation and the improvements required.

1.4.3 Methodological procedure

To fulfill the proposed objectives, a methodological procedure was defined following the flow diagram in Figure 1.1.

The basis of the thesis is the corpus, well defined to fit the objectives required in the thesis. Acoustic and lexical analysis are run on parallel over this corpus to understand the possible distortion existing in the speech production of these special users and the lexical modifications they produce over the canonical transcriptions of the words. These modifications differing of what is considered normal speech (at the acoustic and lexical level) will bring up the need of providing Speaker Dependent (SD) acoustic adaptation and lexical adaptation. These two approaches will be firstly studied separately and then in a joint approach to understand how they interact together to provide higher performance over the baseline ASR system.

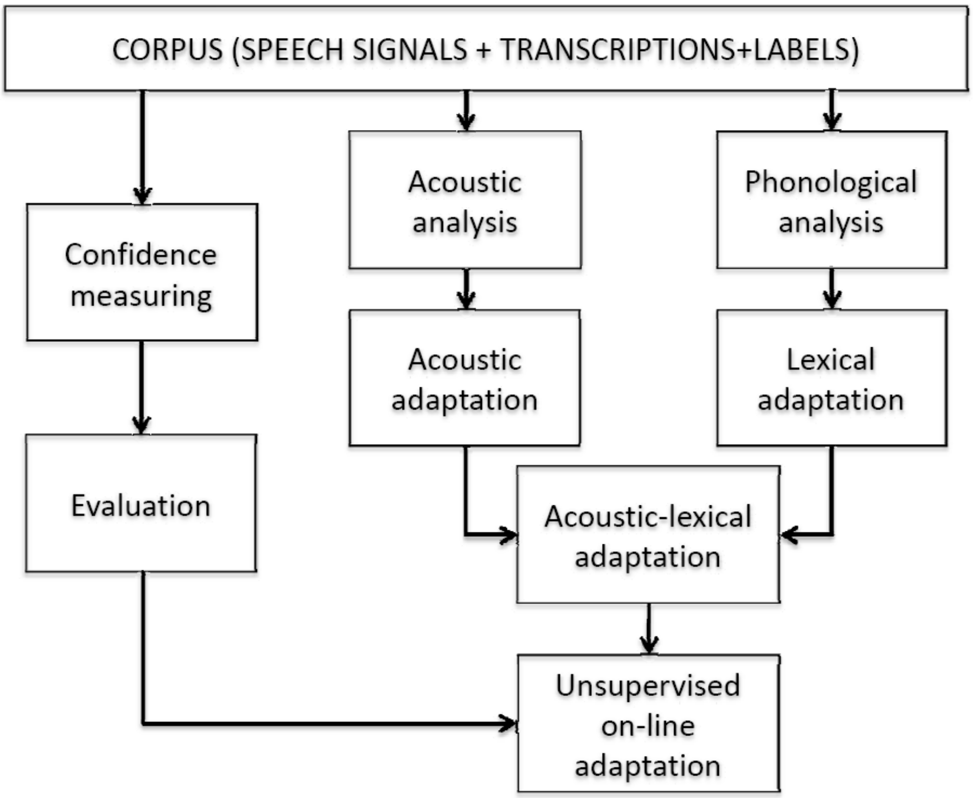

CORPUS (SPEECH SIGNALS + TRANSCRIPTIONS+LABELS)
Confidence measuring
Acoustic analysis
Phonological analysis
Acoustic adaptation
Lexical adaptation
Evaluation
Acoustic-lexical adaptation
Unsupervised on-line adaptation

Figure 1.1: Methodological structure of the thesis

A separate line studying confidence measuring in this corpus to evaluate and detect the mispronunciations and the quality of pronunciation of these speakers will be run at the same time. This confidence measuring will have to achieve a decent performance that makes it able to be introduced in automated systems for speech therapy that can provide an effective feedback in young learners.

Final aim is to put the knowledge in all these areas in an unsupervised adaptive SD ASR system that can provide of sufficiently accurate ASR performance for these users without any prior enrollment stage. Final step will be to advance to the next level of personalization, in which more realistic human-machine interaction will be studied and evaluated.

1.5 Organization

The organization of this thesis, shown in this Section, is intended to provide a straightforward explanation of the work made during it following the methodological procedure of Figure 1.1; starting from the collection of resources, the analysis of them and the scientific advances of the thesis. From this point, the reader can jump to any point of interest according to the brief presentation of the Chapters which is provided.

1.5.1 Corpus and baseline

A deep review in previous efforts for the topics of this thesis is provided in Chapter 2. These efforts comprehend the recording of corpora containing speakers with different types of oral communication disorders, the development of tools for the handicapped (ASR-based systems, Speech Generating Devices (SGDs), ...) and the creation of tools for CALL and Second Language (L2) learning. This review will be shown to be useful for identifying the best ideas to be learned from previous experiences within this kind of work. Much of the work proposed for this thesis came from the open lines left by previous works, from interesting points of discussion shown by other authors or from remarkable results obtained in their work. For all these reasons, the whole Chapter is dedicated to acknowledge and appreciate their extensive work.

Chapter 3 will provide a full overview of the corpus used for the research and results in this thesis. This corpus was obtained with the effort of two separate institutions, the Instituto de Investigación en Ingeniería de Aragón (Aragon Institute for Engineering Research) (I3A) and the CPEE "Alborada", and contains 14 young impaired speakers with different oral communication disorders. After reviewing all the previous efforts in corpora collecting, it was decided to carry on the work of acquiring new data which fitted better the proposals and objectives of this thesis. All the descriptions in terms of speaker and session characterization is fully provided in the Chapter. Further work made over the corpus is also reviewed on it, including a parallel corpus with speech from a large number of unimpaired children and an extensive labeling of the mispronunciations in the disordered data uttered by the impaired speakers.

In Chapter 4 the results of the proposed baseline ASR system over the corpora are provided. This system is a conventional state-of-the-art system, based on Hidden Markov Models (HMMs) and Viterbi decoding. This system will show a fine performance on unimpaired children speech, but a major degradation will be seen when facing disordered speech from the young speakers in the corpus. The influence of task and domain adaptation will be shown, isolating the effect of speech impairments in the task of ASR. Studies on possible topologies (from word to subphone models) are also shown using in the two different approaches evaluated (Task Independent (TI) and Task Dependent (TD) models). Results in the task of Acoustic Phonetic Decoding (APD) will also be reviewed in this Chapter as a possible estimator of phonetic sequences in the corpus; and finally the studies will be completed with an evaluation of the correlation between the error rates (Word Error Rate (WER) and Phoneme Error Rate (PER)) of the systems and the mispronunciations made by the impaired speakers, identifying how different aspects within the disordered speech affect their performance.

1.5.2 Analysis of disordered speech

Chapter 5 will review the influence of the speakers' disorders in the acoustic properties of their recorded spoken utterances. Due to the difficulties of matching the articulatory properties of consonantal sounds to the acoustic field, only vocalic sounds will be used for this study, described in terms of their formant frequencies, pitch frequency, energy and length. The vocalic segments within these utterances will be obtained and the acoustic distortion introduced by the impaired speakers over these signals wil be calculated with respect to the unimpaired reference speakers. The study of these four features will come up with a measure of the acoustic distortion in the disordered speech, that will be especially marked in the loss of discriminative ability between the formant distributions of certain vowels.

Chapter 6 will evaluate how speakers change the lexicon in their pronunciations from the canonical transcription of the uttered words. This evaluation will take as starting point the human labeling on the speaker mispronunciations and will provide a separate mispronunciation rate for each phoneme. Understanding how the speakers are highly consistent in their production at the lexical level will allow for a deeper analysis that will try to know if there are certain patterns in the way in which speakers produce their mispronunciations related to the phonetic context or to the syllable context. Syllable context will appear, then, as an effect with a major impact in the production of mispronunciations by the impaired speakers. The final comparison with the patterns of phonological acquisition in young unimpaired children will allow for hypothesizing the origins of the functional speech disorders in the speakers.

1.5.3 Improvements in ASR and PV

As an outcome of the understanding of the origins and effects of speech disorders in Chapters 5 and 6, Chapter 7 will come up with the need of applying acoustic and lexical adaptation in ASR of disordered speech to neutralize the pernicious effect of the acoustic distortion and lexical variants. First, acoustic adaptation will be reviewed with different possibilities based on the different transcriptions that can be fed to the adaptation framework; later, lexical adaptation with different approaches will be studied, including data-driven and rule-based methods. Finally, a combined acoustic-lexical adaptation framework will be studied that will make use of all the knowledge acquired previously and will try to understand how both techniques interact together in providing a major performance increase in ASR systems.

Chapter 8 will study the possibilities in confidence measuring and scoring systems for assessment of speech quality and pronunciation quality over the impaired speakers. This task will be shown to be similar to the traditional speaker verification task and this fact will allow to determine the best actions to take to improve the system. The suitability of the scores obtained with the acoustic models used for ASR will be studied and it will be seen how all elements of variability in speech (speaker, channel, session...) do not allow for an accurate decision in the detection of mispronunciations. As a first step, traditional and novel score normalization methods will be tested, showing that more accurate phonetic knowledge is useful for this task. These novel systems and methods will be evaluated taking advantage of the special lexical properties of the disordered speech to detect and predict these phonetic mistakes. Speaker adaptation will be used too to battle the effect of speaker variability in the overall detection problem, in parallelism with the speaker verification task, achieving an improved performance in terms on Equal Error Rate (EER).

1.5.4 Personalization of speech-based systems for the speech impaired

Chapter 9 will gather all the knowledge acquired in the previous Chapters and will propose a framework for unsupervised ASR adaptive systems. This will take the acoustic and lexical adaptation studied in Chapter 7 and will transfer it to an unsupervised framework that will make use of the scoring systems studied in Chapter 8. This system will try to provide improvement in the performance within a more realistic approach in developing ASR-based devices for the speech handicapped, because it avoids long and exhausting enrollment sessions that are required in supervised approaches. In this proposal, the system has to adapt itself as the user makes use of the system without any a priori knowledge of the speaker's utterances, which will be seen to be a major problem in the presence of impaired inaccurate speech.

Chapter 10 will make a review of the tools and devices developed through and on parallel to the thesis. Tentative development of speech-based control devices for the handicapped will be commented with different systems like a speech-activated wheelchair or voice-acted keyboard and mouse emulating elements. As major contribution in this field, the set of Computer-Aided Speech and Language Therapy (CASLT) tools developed by the group in the "Comunica" framework will be presented. The four tools which are part of "Comunica" are intended to provide a semi-automated method for speech therapy and they have a strong relation with this thesis. The tools aim to provide all degrees of speech therapy from simple phonation to language abilities like syntax or semantics. Some of the results of the thesis were incorporated in these tools in terms of ASR and

Pronunciation Verification (PV) for evaluation, while extending the tools to Spanish L2 learning, resulting in encouraging experiences in real situations.

1.5.5 Conclusions

It will be in Chapter 11 where a whole discussion of the results in the thesis will be provided. The discussion will focus on trying to understand the origins and effects of the speakers' disorders in their speech, according to the analyses and results of the thesis. Then, the possibility of developing personalized adaptive systems for handicapped individuals will be put under examination according to the results, which will show the real expectancies of these systems. The strong and weak points of the proposed studies will be shown all through the thesis and interactions between different elements studied in the thesis will be also pointed out. Finally, a critical approach to the development of tools for the handicapped will be made according to the experience gained in this work.

Finally, Chapter 12 will provide the conclusions to this work. A small summarization of the work and results will be made, prior to indicating which ones of the initial objectives of the thesis have been fulfilled and which ones could be the next lines to follow in the work in this specific area of research. Further work has arisen during this thesis and some key points of the arising possibilities are opened and proposed, encouraging future researchers to move into this area fearlessly.

1.5.6 Appendices

Appendix A is aimed to provide those readers who are unaccustomed to work in the Spanish language with a small review of all the phonemes and sounds in Spanish, as their distinction is used in many parts of the thesis. Appendix B just presents a small review on adaptation techniques and in the actual implementation of them made for the thesis. Finally, Appendix C will present the confusion matrices for some of the recognition experiments performed through the thesis.

Chapter 2

Review of Previous Resources

Julien knew no Latin except the Bible

-Marie-Henri Beyle 'Stendhal', *The Red and the Black*

As shown in Chapter 1, this thesis aims to advance in the research on the use of speech technologies for the improvement of the quality of life of handicapped individuals through speech-based tools and devices. The variabilities in the speech in these cases are extraordinary as they merge acoustic and lexical variability and only comparable to other situations like non-native or dialectal speech.

This research does not follow the mainstream of the current speech research and, hence, the amount of resources available is significantly smaller than that of traditional tasks in speech technologies like acoustic modeling, robust ASR or speaker verification and recognition which are usually oriented to normal healthy speech. For this reason, this Chapter brings a review of those elements that can be established as an state-of-the-art basis for this research line. This Chapter also serves as a recognition to previous efforts in this area, from which this thesis has inherit a lot of interesting ideas and proposals.

Furthermore, this Chapter intends to be more a review in resources than in technologies or algorithms. Most of the tools and developments reviewed in this Chapter make use of well-known techniques in ASR of assessment adapted to the specific tasks depicted here. A full review of these techniques might make the reader to lose track of the interests of this thesis (speech technologies for the handicapped), so all those references understood as most relevant are provided in this Chapter for further reading on the specific techniques applied in each work.

The Chapter is organized as follows: Section 2.1 will review speech corpora in the area of voice, speech and language pathologies and disorders. The features of these corpora (speakers, amount of data, etc) will be explained and the strong points for research of all of them will be exposed. Posteriorly, Section 2.2 will review different systems based on speech technology for improving the quality of life of the handicapped; these systems make use of ASR or Text-To Speech (TTS), among other technologies. Finally, in Section 2.3, state-of-the-art systems for CALL will be reviewed in different areas like speech therapy for handicapped individuals or L2 learning for foreigners.

2.1 Speech Corpora

Speech data, collected in different corpora covering all possible sources of variability in speech like language, age, gender, acoustic environment and a long etcetera of possibilities are the test bench in which all speech researchers put the foundations of their work. Corpora that are widely accepted by all researchers from different countries and groups allow them to test their algorithms and systems in the same benchmark and, hence, evaluate the improvements they achieve in the different tasks. An important issue in these speech corpora is, apart from a good characterization and organization of the data, to gather the biggest possible amount of speech as possible for a better generalization of the results that are achieved. Hence, massive corpora are being used nowadays for all speech-technology related tasks.

English is, with a remarkable difference, the language that gathers most of the corpora widely used in speech research. The most extended corpora in different areas are the Texas Instruments/Massachusetts Institute of Technology (TIMIT) database [Garofalo et al., 1993] for ASR and APD, Texas Instruments Digits (TIDigits) [Leonard and Doddington, 1993] for isolated-word ASR, International Computer Science Institute (ICSI) Meetings [Janin et al., 2004] for spontaneous speech ASR, Aurora2 for robust ASR in noisy conditions [Hirsch and Pearce, 2000], the National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) or Language Recognition Evaluation (LRE) corpora [Martin and Przybocki, 2004] for speaker and language recognition respectively, or the Speech Under Simulated and Actual Stress (SUSAS) [Hansen, 1999] for speech under stress conditions ASR.

Regarding other languages; more precisely Spanish, which is the target language of this thesis, the most prominent corpora include the Albayzín [Moreno et al., 1993] corpus for ASR or the Spanish subset of the SpeechDat-Car corpus [Moreno et al., 2000] in car noisy environments. For other tasks, further resources are available like the Domolab corpus [Justo et al., 2008] for dialogue systems in a home environment or Audio Visual at Car (AV@CAR) [Ortega et al., 2004] for multi-modal in-car interaction. These corpora contain speech from adult unimpaired speakers in the Iberian peninsula dialect of Castilian-Spanish.

Concerning available corpora that collects disordered speech, a further review is provided over what have been seen as the most relevant efforts up to this date. Sparsity of data is a relevant and, somehow, unavoidable matter in these speech resources: The access to a sufficient number of speakers that deal with the requirements set in the corpus is usually complicated; and, furthermore, recording sessions are a major challenge for these speakers, as their speech processes are tiring and difficult and speakers get exhausted easily.

2.1.1 Disordered speech corpora in English

Two corpora in English language were recorded in the earlier years in which disordered speech was starting to become relevant in the research in speech technologies: The Whitaker database and the Nemours database. These corpora were used for a long time in the initial studies of ASR for disordered speech like [Deller et al., 1991].

The Whitaker database [Deller et al., 1993] contained speech from 6 dysarthric speakers with cerebral palsy. Each speaker repeated 30 times a set of 81 words. An unimpaired control speaker was also recorded, with 15 sessions of all the words. The set of words was made of 46 isolated words (10 digits, 26 alphabet letters, 10 commands) and 35 words from a long passage. The total amount of speech was, hence, 2430 words from each impaired speaker and 1215 words from the unimpaired reference speaker.

The Nemours database [Menéndez-Pidal et al., 1996] gathered speech from 11 speakers suffering dysarthria. Each speaker uttered 74 meaningless sentences, where each sentence was following the structure $The\ X\ is\ Ying\ the\ Z$ with X and Z chosen randomly from a set of 74 monosyllabic nouns and Y from a set of 37 monosyllabic verbs. Furthermore, two whole paragraphs were recorded from each speaker to complete the corpus. All speakers were assessed by a speech pathologist to evaluate the overall quality of their speech and all the corpus was labeled and word and phonemes borders were set within the sentences in the corpus. The main interest of this corpus is how it gathered connected speech from the dysarthric for the evaluation of the difficulties that uttering full sentences created in these speakers.

Concerning pathological speech and voices; usual corpora are recorded containing utterances from normal and pathological voices. Sustained vowels and isolated words are usually recorded in these corpora for the detection of voice abnormalities in the patient's speech as the ones presented in the review in Section 1.3.1. The Disordered Voice Database (DVD) [Disordered Voice Database, 1994] is one of the most accepted databases for the task of assessment of voice pathologies.

Many other corpora have been used for the research in disordered speech; although in many cases they have been recorded as ad-hoc corpora created specifically for a given task and/or language. This way, few databases have managed to become a generalized benchmark for the study in these tasks, limiting the continuity of the work in this area. The large differences between different impairments and their affections over speech have limited also the generalization of the corpora, as each corpus can only cover a small set of this wide range of possible disorders.

In recent years, novel efforts have supposed the acquisition of more recent corpora which aim to fulfill new requirements in the task. Regarding these new corpora, the Universal Access Database (UADatabase) [Kim et al., 2008] contains speech from 19 speakers (14 males and 5 females) whose intelligibility has been assessed by a set of experts, obtaining ratings as low as 6% of intelligibility for some of the speakers. Each speaker uttered 765 isolated words (155 words repeated in 3 different sessions and 300 words with only one repetition) to obtain a total of 14,535 isolated words for the corpus. This UADatabase contained 8 different channels of speech and a video channel of the speakers uttering all the words for the requirements of multi-modal ASR [Sharma and Hasewaga-Johnson, 2009], which has been shown to improve the accuracy of the systems via different techniques like lipreading [Potamianos and Neti, 2001].

2.1.2 Disordered speech corpora in Spanish

In Spanish language, the work in characterization of disordered speech started in more recent dates than in other languages like English; however, a very interesting and well designed corpus can be found in Spanish: the corpus from the Herramienta de Ayuda para la Confidencia de Realizaciones Orales (Help Tool for the Confidence of Oral Utterances) (HACRO) project [Navarro-Mesa et al., 2005], which contained disordered speech as well as unimpaired control speech for the development of pronunciation assessment tools. The distribution of speakers was as defined in Table 2.1 for a total of 43 impaired speakers and 19 unimpaired speakers. Each speaker uttered one session of the 57 words of the speech therapy register RFI [Monfort and Juárez-Sánchez, 1989] which was introduced in Section 1.3.4 as a major resource of speech therapy evaluation in Spanish. Phoneme boundaries were fully identified in the speech signals, characterizing phoneme mispronunciations, deletions or insertions.

Table 2.1: Speakers in the HACRO project corpus

	Impaired Speakers		Unimpaired Speakers	
Age group	Males	Females	Males	Females
6-12 years old	3	2	2	1
12-15 years old	2	3	0	1
15-35 years old	6	2	3	3
35-60 years old	17	3	2	3
>60 years old	2	3	2	2

2.2 Speech Technology-based Systems for the Handicapped

In this Section, different possibilities and developments of applications for the improvement of the quality of life of handicapped individuals will be introduced, with references to relevant research works and projects. Three areas of research are distinguished: Voice assessment, oriented to clinical applications and treatments in cases of voice and speech pathologies; systems based on ASR for the development of speech interfaces for the handicapped; and SGDs, which make use of speech synthesis for creating oral output interfaces to improve communication in cases of severe speech disorders.

Another different set of applications oriented to improve the communication of the handicapped, hearing impaired in this case, have been developed through the years; including cochlear implants or systems for translation from speech to sign language and viceversa [D'Haro et al., 2008]. Although there has been a lot of work involved in this area, they will not be covered by this review, as the concern in the thesis are the aids for individuals for communicative disabilities in their speech and not in their hearing, even if in many cases hearing disabilities produce speech disorders as it was seen in Section 1.3.

2.2.1 Assessment of voice pathologies

Voice pathologies have been shown in Section 1.3 as one of the major causes for impairments in the language. The evaluation and treatment of these voice disorders differs greatly from the cases of speech disorders. As voice pathologies have a physical cause (congenital or acquired), their treatment requires a clinical approach like surgery; speech disorders, on the contrary, are related to the mental process of language and speech and can only be treated from educative approaches. Hence, the review of different CALL tools for the speech training of the handicapped will be made in the next Section; and only the specific developments for voice pathology assessment will be reviewed here.

Two different approaches are usually taken within the task of detecting voice abnormalities in the patient's speech. The first one is based on acoustic features and the second on statistical methods:

Voice analysis based on the study of different speech features can help to detect incorrect voice production in a patient. This can be done by means of a group of acoustic parameters [Yu et al., 2001] in which jitter and shimmer play an important role as jitter is measuring the abnormal short-time variations of pitch while shimmer measures the abnormal intensity changes in the voice envelop [Ruiz et al., 2008]. Other voice analysis approaches use more complex acoustic representation of the pathological voice [Pantazis et al., 2009] or try to obtain a biomechanical model of the patient's vocal tract from the recorded voice of that patient [Gómez et al., 2005], substituting an invasive technique like laryngostroboscopy, for detecting abnormal elements in the vocal tract.

Statistical methods are also used to discriminate between healthy and pathological voices in some cases of dysphonic speech. Traditional techniques that are also used for speaker or language verification like Gaussian Mixture Models (GMMs) [Bocklet et al., 2009] or Neural Networks (NNs) [Godino-Llorente and Gómez-Vilda, 2004] are also used to provide an initial diagnosis on the presence of possible voice pathologies. These methods try to create different statistical models for the healthy and pathological voices in a training phase with labeled data previous to the assessment procedure.

In the end, both types of methods can be complemented to achieve an enhanced performance in the correct assessment of these pathologies, which is an extremely helpful tool for practitioners and otorhinolaryngologists in the diagnosis of vocal chord lesions or head and neck cancer.

2.2.2 ASR for disordered speech

The research in ASR for disordered started in the early 90's with the apparition of the first databases containing this kind of speech. The first remarkable works [Deller et al., 1991] showed up the need of re-thinking the traditional HMM structures to cope with the special features of disordered speech. Since them, dysarthric speech has been the main line of research in the work of developing ASR systems for the handicapped community.

The results of the analytical studies on disordered speech have shown as major conclusions the increase in acoustic variability in the speech production [Blaney and Wilson, 2000], this leading to a loss on intelligibility which relates to the loss of performance in ASR systems [Ferrier et al., 1995]. This larger variability is due to the own nature of these impairments in which the control of the phonation organs is diminished in these speakers [Patel, 2002] due to the several causes of their disorders. Further studies in vowel production by impaired individuals [Croot, 1999, Prizl-Jakovac, 1999] have assured these interpretations of the acoustic distortions in disordered speech [Croot et al., 2000].

This knowledge of the acoustic variability introduced in dysarthric and disordered speech put the basis for further works in the improvement of ASR performance. From this, very different techniques have emerged to face the challenges that each specific type of disorders apart from the big deal of work in adaptive acoustic modeling: The use of phonological contexts and lexicons [Sawhney and Wheeler, 1999] have been proposed in the case of speakers introducing mispronunciations in their speech. The possibility of audiovisual information [Potamianos and Neti, 2001] with lipreading has also been evaluated as a method for improved recognition. Furthermore, as more and more work was being done in this line of research, other languages different than English also added their efforts with initial studies like Dutch [Sanders et al., 2002] and little by little more and more research groups joined these efforts.

Two very interesting projects deserve a little more space in this review: The *Vocal Joystick* [Bilmes et al., 2006] and the *Speech Training And Recognition for Dysarthic Users of aSistive Technology (STARDUST)* project [Hawley et al., 2003]:

The *Vocal Joystick* allows that patients with severe physical impairments like cerebral palsy can control computer devices like a mouse or handle wheelchairs or prosthetic devices by speech. The system operates like a joystick in which each direction of the space is operated by a different vowel from the English vowel map. The device has been thoroughly tested in different situations [Harada et al., 2008] and has proved to be one of the most relevant oral help devices for assistive technology.

STARDUST aimed to create a fully functional system for the control of a home environment for patients with cerebral palsy and different levels of dysarthria based on the speech recognition of a short list of oral commands. With this system, the impaired community could have an easier way to acquire more independence in their life, improving their quality of life [Hawley et al., 2005]. STARDUST tried to deal with all the difficulties in the development of ASR for disordered speech, like the lack of sufficient data for the training of personalized systems [Green et al., 2003] or the acoustic variability in dysarthric speech [Wan and Carmichael, 2005].

2.2.3 Speech Generating Devices (SGDs)

SGDs are systems that improve the oral communication of a heavily language impaired individual by means of digital speech substituting the extremely degraded speech from the user. Systems making use of pre-recorded speech are very usual in the world of Augmentative and Alternative Communication (AAC), as different communication boards make use of digital speech to reinforce the acquisition and use of graphic symbols in impaired persons.

However, TTS systems can provide a more flexible aid as they do not force the user to have a limited number of possible words or sentences to utter. In any case, current TTS systems' have to struggle to raise the rates of naturalness and intelligibility of their voices, which is strongly necessary in SGDs oriented to impaired users. These users, due to their hearing or development impairments are more sensitive to the loss of quality of synthesized speech and several studies have been carried out to detect how their comprehension is affected between natural and synthetic speech. These studies, closer to the psychological and neurological effect of synthetic speech than to the technological development behind it, have shown that impaired individuals have a major difficulty in comprehension of synthetic speech compared to unimpaired individuals of their own age [Koul, 2003], but this can be corrected with the continuous exposing to synthesized speech [Koul and Clapsaddle, 2003]. The difference between impaired and unimpaired individuals grows when the task consists in understanding connected speech like full sentences [Koul and Hanners, 1997] due to the major difficulty of obtaining the message from a signal acoustically different to normal speech like that of synthesized speech.

Traditional SGDs are based on a communication board where the selection of different words or symbols synthesizes the desired sentence to augment the communication possibilities of the subject. However, these persons might feel that they are losing part of their personality for not being able to use their own speech. Because of this, the possibility of re-synthesizing the own speech from the user to an speech with better intelligibility, while keeping the original features from the speaker has become an important line of research recently with several approaches that aim to synthesize the speech of a dysarthric individual erasing the sources of acoustic variability and, thus, augmenting intelligibility [Hosom et al., 2003, Kain et al., 2004].

The final outcome of all the technologies presented up to this moment is the realization of a Voice-Input Voice-Output system, as depicted in Figure 2.1. This system processes the oral input of an impaired speaker with a major speech difficulty, performs ASR decoding to obtain the uttered sentence or words and re-synthesizes the speech in a more intelligible voice. The system may also analyze the speech from the user to extract a set of features that allow the new voice to sound similar to the original voice of the user, with the methods presented before.

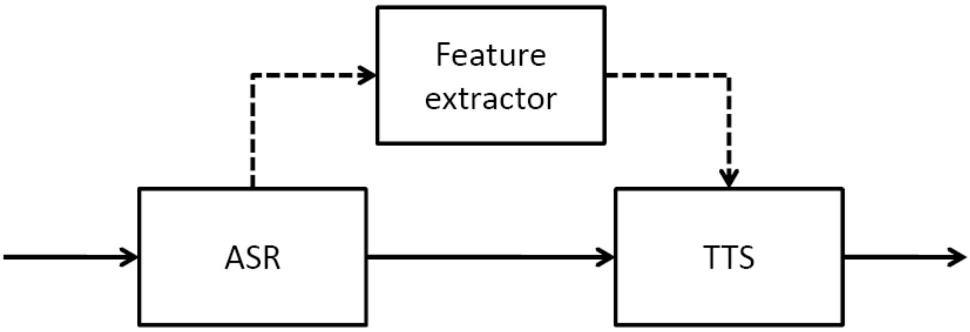


Figure 2.1: Scheme of a Voice-Input Voice-Output system

Although this approach has been evaluated with Linear Prediction Coefficients (LPC)-based and corpus-based synthesis, these techniques are taking advantage recently of the possibility of HMM-based synthesis to provide an enhanced synthetic voice with a closer similarity to the user's voice.

This approach, merging ASR and TTS, had the Voice-Input Voice-Output Communication Aids (VIVOCA) project, which took as starting point the knowledge in dysarthric ASR achieved in the STARDUST project and aimed to create a full communication system, where the user's speech could be recognized and re-synthesized via an SGD [Hawley et al., 2007, Creer et al., 2009, Creer et al., 2010]. The final intention of the project was to be able to use all this technology within a portable device like a Personal Digital Assistant (PDA) which could accompany at any time the impaired individual.

2.3 Computer-Aided Language Learning (CALL) Tools

In this Section, different approaches to CALL systems developed during the recent years will be reviewed. A traditional separation of CALL tools is shown in Figure 2.2, where two dimensions are shown: The vertical dimension defines the end user of the application (CASLT tools are oriented to native speakers with language impairments vs L2 tools for non-native learners) and the horizontal dimension defines which language ability is being trained by the tool (Computer-Aided Pronunciation Training (CAPT) for phonological training vs other abilities like tutors for vocabulary, syntax, grammar,etc).

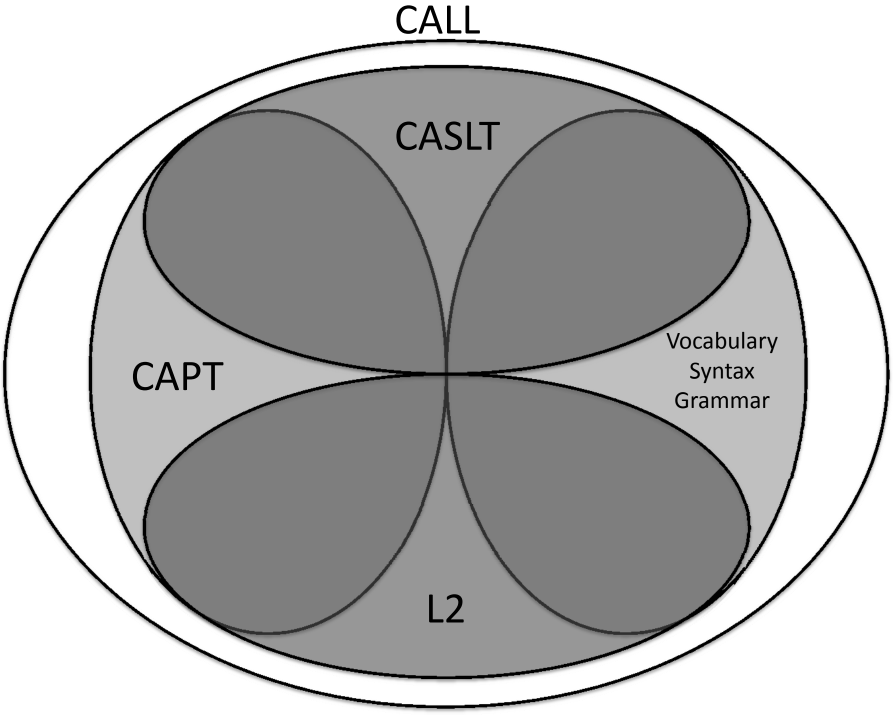


Figure 2.2: Organization of CALL tools

The focus of the review in this Section will be on oral-input based CAPT tools, where the main objective is the objective assessment of the quality of the user's pronunciation, although there is also a big deal of work oriented to provide feedback on the grammatical and syntax abilities of the student. As this thesis is focused on disorders on the speech production, those tools like Reader-Specific Lexical Practice for Improved Reading Comprehension (REAP) [Pino and Eskenazi, 2009, Marujo et al., 2009] or many others which aim to improve the vocabulary or reading comprehension of the student purely by text will not be reviewed.

All CALL tools also require to take care of the different educative aspects of their design, like presenting in an appropriated way the feedback or providing an stimulating environment to the student.

2.3.1 Computer-Aided Speech and Language Therapy (CASLT) tools

CASLT tools are oriented to the improvement of the speech quality in speakers with speech disabilities and to the improvement of their linguistic abilities. These tools aim to provide a help to traditional speech therapy techniques shown in Section 1.3.4 in a semi-automated way, supporting the work of the professional speech therapist.

There has been a great deal of relevance given to these tools from the public institutions [Cucchiarini et al., 2008a] in the recent years. The 5th Framework Program (FP) of the European Union (EU) "Quality of Life and Management of Live Resources" covered several projects for the development of tools like Ortho-Logo-Paedia (OLP) [Oester et al., 2002], SPEech COrrector (SPECO) [Vicsi et al., 1999] or ISAEUS [García-Gómez et al., 1999].

There are three major points to study in all systems to understand how they work [Saz et al., ress]: Target users, selection of the interface and the provided feedback.

The *target user of a speech therapy tool* has to be defined prior to the development of the tool. Different systems are aimed to cover different groups of students or different educative needs. Tools like the Technology-based assessment of language and literacy (Tball) project in [Black et al., 2008] are oriented to correct difficulties in phoneme acquisition in young children in a pre-linguistic phase. In the case of reading tutors, the target users are young adults who need to train their language expression and comprehension. SPECO [Vicsi et al., 1999] performed pronunciation training for users of different European Union countries like Hungary, Sweden or United Kingdom, focusing on the requirements for the development of a multi-language tool. A very specific set of tools are oriented to the hearing impaired community like [García-Gómez et al., 1999, Lefévre, 1996] and their acquired pronunciation difficulties.

Developing an *adequate interface for the target speakers* is important, with the use of AAC elements considered as a strongly requirement for some research systems [Granstroem, 2005]. The interface has to be well matched to the target users, as it is not the same to develop a tool for small children (who will need a more motivational interface) than adults (which require a more precise feedback), as well as considering possible disorders of the patient: hearing or visual impairments might limit the interface possibilities of some users and development disabilities require of an interface that help these users to focus on the educational activities of the tool.

Finally, providing the *correct type of feedback* is necessary to make the tools useful for the speakers. An articulatory feedback can be provided to the user, in which the system is correcting the position of the elements of the speaker's vocal tract like in OLP [Hatzis et al., 2003, Oester et al., 2003] and its predecessor Optical-Logo Therapy (OLT) [Hatzis et al., 1997, Hatzis, 1999, Hatzis et al., 1999]. Purely acoustic feedback as a scoring measure of the quality of the user's pronunciation is a most extended way of feedback in many other systems [Tepperman et al., 2006, Duchateau et al., 2007], based on quality measures similar to the Goodness Of Pronunciation (GOP) in [Witt and Young, 1997]. Finally, reading proficiency and text comprehension trained in reading tutors are measured in different ways: The number of words per minute or the rate of disfluencies measure the children' reading ability [Cleuren et al., 2006], and the comprehension of a given text is measure within a dialog with the student [Gerosa and Narayanan, 2008] in which several questions about a given text have to be answered.

2.3.2 Second Language (L2) learning tools

One of the most successful areas of CALL systems nowadays is the training of students in a foreign language with L2 tools. The need for improving the speaking ability in a foreign language in the global world of today has produced a boost in the development of these tools in both research and commercial areas.

One of the major difficulties for this task is the lack of normalized, well-established corpora due to the large divergence of interests of each group, as this task depends on the L2 to teach and the original First Language or mother tongue (L1) from the speaker. As English has been one of the most researched languages for L2 training, corpora including speech from non-native speakers in English have been gathered containing speakers from Asian countries or different Europen countries like the Italian and German Spoken Learner's English (ISLE) corpus [Atwell et al., 2003]. The main interest of all L2 corpora is to gather speech from speakers uttering words and sentences in a non-native language in which they may have different levels of knowledge.

As mentioned, early L2 tools were oriented to the learning of English by foreign speakers, although nowadays a large number of languages are being covered by L2 tools, turning this line of research as one of the most noticeable in the development of CALL devices. Other languages with remarkable efforts in the research in L2 have been Japanese [Tsurutani et al., 2006], Dutch [Neri et al., 2006] or Chinese [Luo et al., 2008]. More novel languages like Polish [Cylwik et al., 2009] are developing into this area through the EURONOUNCE project and corpus, as well as Norwegian [Amdal et al., 2009], Swedish [Wik et al., 2009] and many others.

Regarding the specific linguistic abilities that these tools aim to improve, a big deal of research is put on the development of CAPT tools for L2 training. These tools, hence, try to fulfill the possible lack of pronunciation training in traditional L2 learning schools where a lot of relevance is given to the teaching of vocabulary, grammar and syntax, neglecting the pronunciation skills of the students [Morley, 1994]. In this CAPT approach, the L2 tools ask for the student to utter single words or sentences and evaluates the phonetic accuracy of the pronunciation [Mak et al., 2003, Chou, 2005, Neri et al., 2006, Cucchiarini et al., 2007, Luo et al., 2008, Fengpei et al., 2008] with different techniques and measures like Utterance Verification (UV) [van Doremalen et al., 2009] or the well-known GOP [Witt and Young, 1997, Kanters et al., 2009].

More novel tools are being researched nowadays towards the elimination of the foreign accent in non-native speakers via the study of pitch and duration cues [Amdal et al., 2009, Cylwik et al., 2009, Szaszák et al., 2009]. Also, higher linguistic capabilities like grammar [Lee and Seneff, 2006] and vocabulary [McGraw et al., 2009] are target elements of these tools. In the end, the gathering of all these different tools might be used for the creation of full courses in the new language [Cucchiarini et al., 2008b] covering all the levels of language in an unsupervised or semi-supervised way.

Although L2 teaching was not initially an objective of this thesis, it was seen during during this work that the research in CAPT for this task could be extrapolated to CASLT, as long as the special needs of patients of speech therapy were had into account in the terms explained in the previous Section (mainly interface and feedback suitable for the handicapped). For this reason, this Section has brought a small review of the current research, focusing in the points of coincidence with possible CASLT tools.

Chapter 3

Alborada-I3A Corpus

"Whenever you feel like criticizing any one", he told me, "just remember that all the people in this world haven't had the advantages that you've had."

-Francis Scott Fitzgerald, *The Great Gatsby*

This Chapter presents the corpora used for all the analyses and experiments carried out on children' speech and disordered speech in this thesis. After reviewing in Chapter 2 all the available resources, collected by previous researchers in the recent years, the motivation for exploring new possibilities in data acquisition appeared, trying to fill the gaps in the recordings acquisition covered by previous works.

The corpus was given the name of "Alborada-I3A" [Saz et al., 2008a], since these two institutions were involved in the speech acquisition process. The I3A supported this thesis and gave all the required framework for the research work during this time; and the CPEE "Alborada" was the main connection with the world of special education providing the facilities and speakers for the speech collection.

The chapter is organized as follows: In Section 3.1, the motivation and objectives for the collection of the speech data will be provided. Section 3.2 will present the environment in which all the recordings were made. The selection of speakers and their characteristics will be presented in Section 3.3, and the organization of the recording sessions in Section 3.4. Finally, the two tasks added at the end of the recording process will be presented: Section 3.5 will describe a parallel corpus containing reference speech from children and young adults; and Section 3.6 will point out the results of a human labeling for the detection of mispronunciations in the corpus of disordered speech. An appreciation to all the people that made possible the creation of these corpora will be given in Section 3.7.

3.1 Motivation and Requirements for the Acquisition

Before starting the acquisition of a new corpus for speech research, it is strongly necessary to evaluate the real need and convenience of it. The process of data collection can be really time demanding for the people in charge of the recordings and for the subjects who donate their speech. Hence, a deep process of evaluation prior to the speech acquisition was required to make sure that the collected speech was really useful for the research purposes it was intended to fulfill.

In the case of the corpus that is to be presented in this Chapter, the motivation for the speech acquisition arose after the review of the previous speech corpora studied in Section 2.1. The lack of corpora for this task in Spanish is only avoided by the corpus of the HACRO project [Navarro-Mesa et al., 2005] seen in Section 2.1.2. The review over this corpus showed that it was a very interesting corpus as it contained speech from speakers with different speech impairments; but the lack of sufficient data to characterize each speaker made it unable to apply supervised or unsupervised adaptation techniques like it has been shown as an objective of this thesis.

Several requirements were set prior to start the speech acquisition, concerning the selection of the speakers and the speech acquisition process. All of them are listed here:

- The acquired speech signal had to assure good quality in terms of low presence of noise (both additive and convolutional noises), to avoid its impact in ASR performance. No methods for noise robustness will be studied in this thesis.
- The speech from all speakers had to be natural and the influence of external elements had to be minimized. Realistic speech will make easier the translation of the results in the simulations to real world situations.
- There had to be a balance in terms of gender among speakers. All the results achieved had to be gender independent.
- The recorded speakers had to be balanced also in terms of age (within the availability in the CPEE "Alborada", where students can only attend classes until they become 21 years old) to also obtain age-independence.
- Furthermore, speakers had to be balanced in terms of the degree of their impairments; this is, containing heavily impaired individuals as well as persons with slighter disorders in their speech. Relationship between degree of disorder and performance was seen as an important effect for study.
- Several sessions had to be collected from each speaker, separating different sessions in different days to reflect intra-speaker variability. This also assures realistic speech, closer to real situations.

Finally, the speech had to be donated freely by the speakers, or with the consent of the parents or tutors when they could not take the legal responsibility for it. Due to this donation, the speech could not be used for the commercial development of systems in which a monetary profit could be obtained. This way, the speech corpora presented in this Chapter is totally open and available for the distribution among research institutions that accept the conditions imposed by this agreement: to use it only for research purposes and to keep the confidentiality of the speakers' data as they are covered by laws protecting the rights of children and handicapped individuals.

3.2 Recording Environment

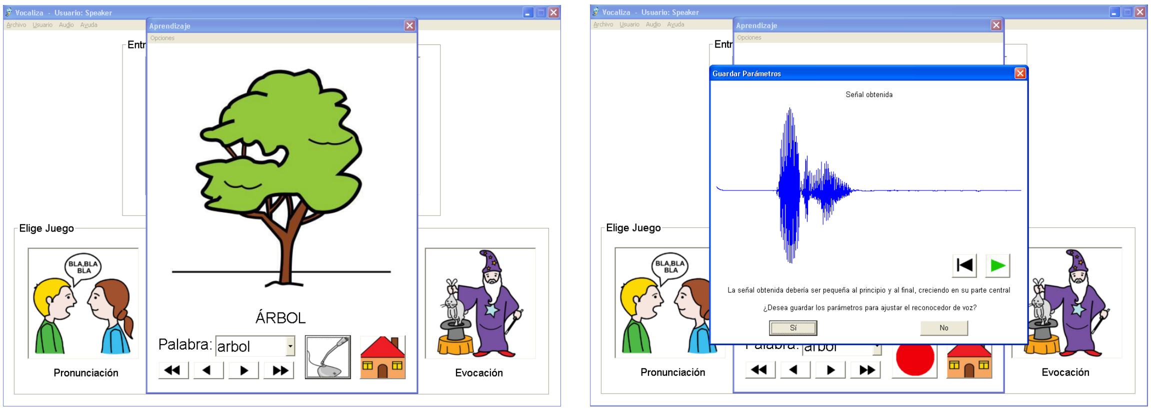
The environment of the recordings was chosen to fulfill two main goals: To achieve the best quality in the recorded speech signals while assuring the comfortability of the speakers to obtain the most natural speech that was possible.

Initially, these two objectives seemed to be divergent. A good quality speech signal is that one which is not corrupted by any kind of noises (additive noise or convolutional noise), but these ideal conditions can only be achieved in an anechoic room within a speech laboratory or in a well-prepared room. However, the target speakers of the corpus are children and young adults with cognitive and social disorders who are very sensitive to new and different environments. Moving them to a laboratory or placing them in a specially prepared room might produce an stress in the speakers which would result in a non-natural speech or even in a total blockage in their language.

Hence, these limitations imposed the physical environment for the recordings: The CPEE “Alborada”. As all the speakers were provided by this educative institution where they attended regular classes, the same classrooms where they were having their classes were chosen for the recordings. For every recording session, an empty available classroom was selected by the staff of the CPEE “Alborada” and the recording procedure was installed (computer, microphone, etc). The target speaker and the person of the I3A who was in charge of the recording each day were the only ones in the room (sometimes a member of the CPEE “Alborada” also witnessed and helped in the process). This way, with the use of a good quality microphone set, external noises could be reduced, but not totally erased because classes were running normally in the rest of the school (with the expected noisiness of the rest of the pupils in the school).

Another element of relevance was the recording interface that the speakers had to face for the recordings. Usual interfaces where the target speaker reads aloud a sentence shown plainly on a screen are good for the acquisition of speech in adults, but it is easy for children (especially, impaired children) to lose motivation with this kind of interfaces. The solution was “Vocaliza”, the speech therapy tool that will be presented in Section 10.2; this tool contained an enrollment phase previous to the speech therapy activities very helpful for the speech acquisition purposes.

With “Vocaliza”, it was possible to create different user profiles and attach some words and sentences to the users that took part in the enrollment phase. During the enrollment, the word or sentence was shown to the speaker via a pictograph or group of pictographs and, optionally, text and synthesized or pre-recorded audio as shown on Figure 3.1(a). For the corpus acquisition purposes, only the pictograph(s) and text were shown to the speaker, and oral reinforcement was directly given by the person of the I3A on duty during the recording process. The person responsible for the recordings navigated through the previously selected words and initiated the recording. When the recording of a word finished, a plot with the signal waveform was shown as in Figure 3.1(b), being possible to listen to the signal as many times as needed to finally accept or discard the utterance depending on the quality of the signal (naturalness, background noise level, etc).



La señal obtenida debería ser pequeña al principio y al final, creciendo en su parte central

¿Desea guardar los parámetros para ajustar el reconocedor de voz?

(a) Word prompting in the corpus recording process

(b) Waveform plotting in the recording process

Figure 3.1: “Vocaliza” interfaces in the speech acquisition

Regarding the hardware that was used for the recordings, the application was running in a commercial laptop with an integrated commercial sound card. More complicated hardware for speech acquisition would have been difficult to handle in the school facilities and uncomfortable for the speakers. In each session, it was checked that no external or internal elements in the environment or in the laptop could create any electromagnetic interferences that might distort or corrupt the recordings with electrical noise.

The final element to take care of in the selection of the recording environment was the microphone. Close-talk microphones assured the best quality in reducing external noise as well as reverberation effects. Unfortunately, they created a feeling of attachment of the speaker to the computer that collided with the objective of making the target speaker feel free and natural during the recordings. Given this, a portable wireless microphone was used so the speakers did not feel totally attached to the computer and, hence, uncomfortable. The final model of microphone chosen was the AKG C444L, which had all the commented features.

With this, all the issues in the environment for the recording have been explained. A graphical illustration of an actual recording situation is shown on Figure 3.2 where all the elements, including computer and microphone, can be seen together.

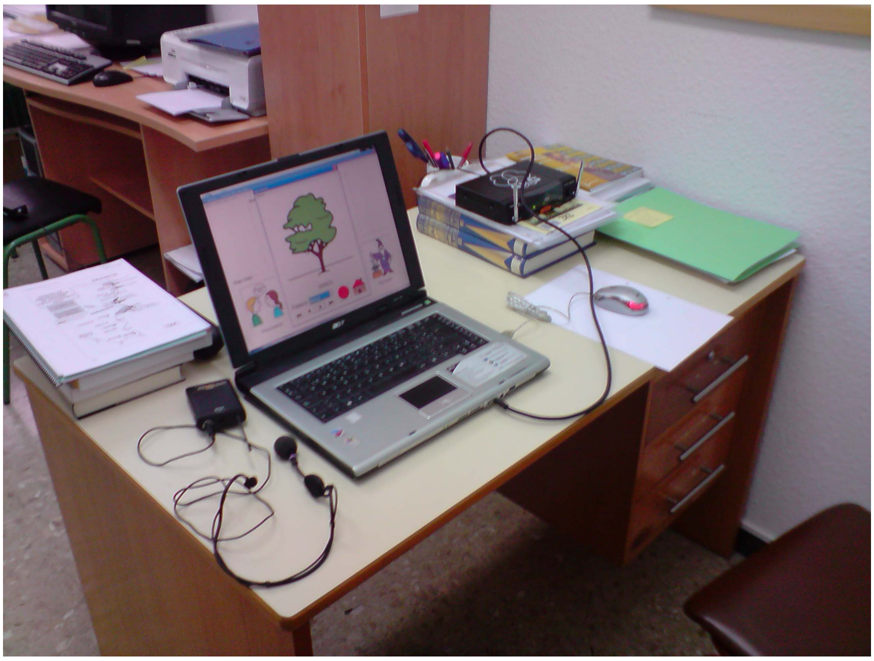


Figure 3.2: Disposition for real environment recordings in the IES "Félix de Azara"

3.3 Speaker Characterization

The process of selection of the speakers was initiated by requirement of the I3A to the staff of the CPEE "Alborada", which provided all speakers to the corpus from their pupils. An initial list of possible speakers was made by them, selecting those whose cognitive abilities could allow them to complete the task of uttering the designed sessions; as subjects with a very high degree of cognitive disorders might have no language at all or might be unable to follow the process of recording. The final selection of speakers was made according to the requirements imposed to the speakers in Section 3.1 in terms of age, gender and disability.

The speaker characteristics in terms of age and gender can be seen in Table 3.1. As it can be appreciated, gender balance was kept as there were 7 boys and 7 girls in the corpus. Age balance was also kept in the range from 11 to 21 years old (histogram is shown in Figure 3.3).

The limitations to the ages of the speakers (range 11-21 years old) were imposed by two factors: Lower boundary (11 years old) was forced by the fact that no one of the children under 11 years was seen to be able to finish the session acquisition process successfully due to their young age; and the upper boundary (21 years old) was forced because this was the maximum age in which they could stay into a public educational institution like the CPEE "Alborada".

Table 3.1: Speakers in the disordered speech corpus

Speaker	Age	Gender	Speaker	Age	Gender
*Spk*01	14 years old	Female	*Spk*02	11 years old	Male
*Spk*03	21 years old	Male	*Spk*04	21 years old	Female
*Spk*05	18 years old	Male	*Spk*06	17 years old	Male
*Spk*07	18 years old	Male	*Spk*08	19 years old	Male
*Spk*09	11 years old	Female	*Spk*10	15 years old	Female
*Spk*11	20 years old	Female	*Spk*12	18 years old	Male
*Spk*13	13 years old	Female	*Spk*14	11 years old	Female

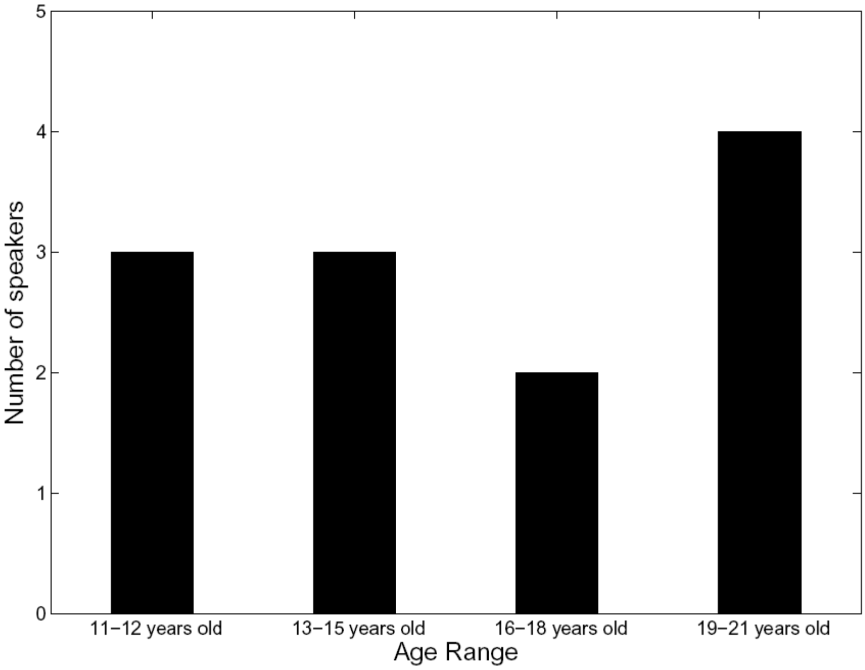


Figure 3.3: Histogram of speakers' ages in the corpus

3.3.1 Speaker disorders

Characterization of the speakers disorders is also necessary in the description of the kind of corpus described in this Section. The speakers may have physical and cognitive disorders that may interfere their speech and language in different ways. A summarized diagnosis for all of them, with special attention to their speech and language disorders is provided in Table 3.2. Each speaker can be, hence, characterized in terms of the different voice, speech and language disorders presented in Section 1.3.

Moderate to severe cognitive disorders were common in all speakers and this producesd different language disorders in their speech. Dyslalias and other speech disorders like dysphemia or some degree of dysarthria were also shared in different amounts by all of them. Speakers suffering Down's Syndrome also suffered in many cases some minor malformations in organs of the vocal tract that could produce a certain degree of dysglossia.

At this point started one of the major differences with previous corpora in this area. This novel corpora focused only in children and young adults and their specificities in speech and language disorders. The impossibility to cover in a properly balanced way a wider range of ages and the availability of these speakers at the CPEE "Alborada" produced this. Also, as the development of speech therapy tools oriented to disabled children (as it will be seen on Section 10) was an important objective during all the thesis, this corpus fulfilled the requirements to study this speech and improve the performance of these tools.

Table 3.2: Disorders of the speakers in the disordered speech corpus

Speaker	Physical and Cognitive Disorders	Speech and Language Disorders
*Spk*01	Down's syndrome	Semantic-pragmatic disorder
*Spk*02	Attention deficit, hyperactivity	Syntax and semantic disorders
*Spk*03	Emotional deprivation disorder	Dyslalia and semantic-pragmatic disorder
*Spk*04	Emotional deprivation disorder	Semantic-pragmatic Disorder
*Spk*05	Down's syndrome	Dysphemia and semantic disorders
*Spk*06	Motor ataxia and tetraplegia	Semantic-pragmatic disorder
*Spk*07	Polymalformation syndrome	Dyslalia and semantic-pragmatic disorder
*Spk*08	Cerebral palsy	Phonological and semantic disorders
*Spk*09	Development disorder	Dyslalia and Specific Language Impairment
*Spk*10	Hypoxic-ischemic encephalopathy	Phonological and semantic disorders
*Spk*11	Cognitive disorder	Semantic-pragmatic disorder
*Spk*12	Development and expressive disability	Specific Language Impairment
*Spk*13	Down's syndrome	Phonological and semantic disorders
*Spk*14	Development disorder	Dyslalia and Specific Language Impairment

3.4 Session Characterization

Three different types of sessions were designed for the impaired speakers chosen for the recordings. The core of the recordings were isolated words, while some sessions with simple and complex sentences were furtherer recorded from some speakers.

3.4.1 Isolated word sessions

Each speaker in the Alborada-I3A corpus recorded 4 sessions with isolated words. The sessions were recorded in different days to obtain a set of sessions which could reflect properly the possible effects of intra-speaker variability. With these 4 sessions per speaker, the corpus contained a total of 228 isolated-word utterances per speaker and 3192 utterances as a whole, being a total of 2 hours, 17 minutes and 4 seconds of speech signal (including silence). The average Signal-to-Noise Ratio (SNR) per utterance was 34.58 dBs with an standard deviation of 6.66 dBs, which reassured the quality of the speech collection carried out during the acquisition process.

The vocabulary for these sessions was the set of 57 words included in the RFI [Monfort and Juárez-Sánchez, 1989]. As seen in Section 1.3.4, this is a well-known handbook for speech therapy in Spanish. Furthermore, it was also the vocabulary for the oral corpus in the HACRO project [Navarro-Mesa et al., 2005] shown in Section 2.1.2. This 57 words and their transcription according to the International Phonetic Alphabet (IPA) and the Speech Assessment Methods Phonetic Alphabet (SAMPA) alphabets are shown on Table 3.3.

RFI contains all the 24 phonemes described traditionally in the Spanish language [Alarcos, 1950] as well as the most usual allophones. The whole description of Spanish phonemes and sounds can be seen in Appendix A, altogether with their IPA and SAMPA transcriptions. Furthermore, RFI contains a major selection of elements of the language, ranging from monosyllabic words to polysyllabic words, with clusters of consonants, codas and diphthongs.

The number of syllables in the 57 words is 129 (2,26 syllables average per word) and the number of phonemes is 292 (5.13 phonemes per word).

Table 3.3: Words in the RFI and their IPA and SAMPA transcriptions
Set of words and transcriptions in the corpus

Word	IPA	SAMPA	Word	IPA	SAMPA
árbol	[’ɑɾβ̞ɔl]	[“ArBOl]	boca	[’boka]	[“boka]
bruja	[’bɾuxa]	[“bruxa]	cabra	[’kaβ̞ɾa]	[“caBra]
campana	[kɑm’panã]	[kAm“pana˜]	caramelo	[kaɾa’melo]	[kara“melo]
casa	[’kasa]	[“kasa]	clavo	[’klaβ̞o]	[“klaBo]
cuchara	[ku’t͡ʃaɾa]	[ku“tSara]	dedo	[’d̪eđ̞o]	[“deDo]
ducha	[’d̪ut͡ʃa]	[“dutSa]	escoba	[es̬’koβ̞a]	[es“koBa]
flan	[’flɑn]	[“flAn]	fresa	[’fɾesa]	[“fresa]
fuma	[’fuma]	[“fuma]	gafas	[’gafɑs]	[“gafAs]
globo	[’gloβ̞o]	[“gloBo]	gorro (cap)	[’goro]	[“gorro]
grifo	[’gɾifo]	[“grifo]	indio (indian)	[’ind̪jo]	[“indjo]
jarra	[’xara]	[“xarra]	jaula	[’xɑwla]	[“xAwla]
lápiz	[’lapi̞θ]	[“lapIT]	lavadora	[laβ̞a’đ̞oɾa]	[laBa“Dora]
luna	[’luna]	[“luna]	llave	[’ʎaβ̞e]	[“LaBe]
mariposa	[maɾi’posa]	[mari“posa]	moto	[’moto]	[“moto]
niño	[’nĩɲo]	[“ni˜Jo]	ojo	[’ɔxo]	[“Oxo]
pala	[’pala]	[“pala]	palmera	[pɑl’meɾa]	[pAl“mera]
pan	[’pɔn]	[“pAn]	peine	[’pɛjne]	[“pEjne]
periódico	[pe’ɾjođ̞iko]	[pe“rjoDiko]	pez	[’peθ]	[“peT]
piano	[’pjano]	[“pjano]	pie	[’pje]	[“pje]
piña	[’piɲa]	[“piJa]	pistola	[pi̞s̬’tola]	[pIs“tola]
plátano	[’platanõ]	[pla“tano˜]	playa	[’plaj̞a]	[“plajja]
preso	[’pɾeso]	[“preso]	pueblo	[’pweβ̞lo]	[“pueBlo]
puerta	[’pwɛɾta]	[“pwerta]	ratón	[ra’tɔn]	[ra“ton]
semáforo	[se’mafoɾo]	[se“maforo]	silla	[’siʎa]	[“siLa]
sol	[’sɔl]	[“sOl]	tambor	[tɑm’β̞ɔɾ]	[tAm“BOr]
taza	[’taθa]	[“taTa]	teléfono	[te’lefono]	[te“lefono]
toalla	[to’aʎa]	[to“aLa]	toro	[’toɾo]	[“toro]
tortuga	[tɔɾ’tuɣ̞a]	[tOr“tuGa]	tren	’[tɾen]	[“tren]
zapato	[θa’pato]	[Ta“pato]			

These isolated word recordings were expanded for speakers *Spk*07 and *Spk*08 who, two years after the first acquisitions, were recorded again with 4 sessions of the words in the RFI. These extra sessions were not used during the main work of this thesis work to avoid the unbalance of the speakers, but it will be seen how they were useful regarding studies on the amount of adaptation data. Speakers *Spk*07 and *Spk*08 were 18 and 19 years old respectively during the initial recordings, so it could be easily argued that their speech characteristics did not change much in these two years until becoming 20 and 21, as most of the speech changes in young adults happens at an earlier age.

An important issue to study when preparing tasks for ASR is to evaluate how distant the words in the vocabulary were from each other. Phonetic distance between all 57 words in the RFI was measured and the histogram is presented in Figure 3.4. Lowest phonetic distance was 2 in 35 cases (for instance in ‘cabra’ vs. ’casa’) and the highest distance was 10 in the case of ‘casa’ vs. ‘periódico’. The average phonetic distance was 4.99 in a total of 1596 possible word comparisons. The phonetic distance was measured as the total number of phoneme substitutions, insertions and deletions required to transform a word into one another. This implied that the half of the phonetic distance between two words was the minimum number of phonetic changes that might produce total confusion between these two words.

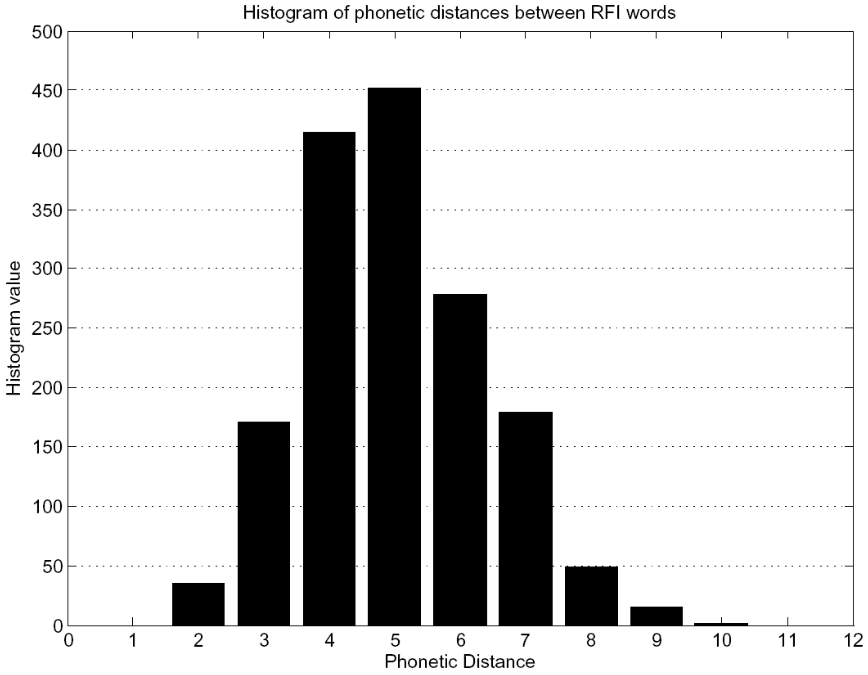


Figure 3.4: Phonetic distance histogram

Example 3.4.1 A simple example of the measure of the phonetic distance can be given between words in the RFI: Words “taza” and “zapato” have a phonetic distance of 4, because it is necessary to make 4 phonetic changes to convert “taza” into “zapato” (substitute [t] for [θ] at the beginning and [θ] for [p] in the middle, and finally insert [to] at the end). Hence, with only two changes it is possible to make both words fully confusable, for instance [taθato] is 2 phonemes away from each one of them; which would make not possible to decode the original word.

3.4.2 Simple sentence sessions

Some of the speakers with better cognitive and language capabilities were also enrolled in a task consisting in short meaningless sentences constructed with words from the RFI. The structure of these sentences was created as follows:

el/la Word1 y el/la Word2

Where $Word1$ and $Word2$ were chosen randomly from the RFI. Connectors *el* and *la* were the determinants (*the* in English) and connector *y* was the copulative joint (*and* in English). This approach resembled the speech acquisition in the Nemours database [Menéndez-Pidal et al., 1996] in Section 2.1, which was seen as an effective way of collecting connected speech from impaired speakers.

Example 3.4.2 Examples of the simple sentences generated for the impaired speakers are:

- la *lavadora* y la *playa*
- el *sol* y el *peine*
- el *ojo* y el *árbol*

Four sessions were designed, with each session containing 28 sentences and each word of the RFI appearing only once in every session (hence, words appeared four times through all sessions). A further restriction was imposed; that in the appearances of each word, two were in the position of $Word1$ and two in the position of $Word2$.

Only 4 speakers, *Speaker*01, *Speaker*04, *Speaker*06 and *Speaker*11, could fulfill these sessions, because the mid to severe range of the cognitive and language impairments of the other speakers made them unable to utter fluently connected speech as the proposed in the sessions. Each speaker uttered, hence, 112 of these simple sentences; and the corpus contained 448 of them, for a total of 25 minutes and 30 seconds of speech including silence. The mean SNR per utterance was 34.05 dBs with an standard deviation of 6.72 dBs.

3.4.3 Complex sentence sessions

Finally, three speakers: *Speaker*04, *Speaker*06 and *Speaker*11 were also enrolled in a short session with 10 full meaningful sentences. Restrictions over these sentences were that they had to contain three different words of the RFI, and the total number of words had to be lesser or equal than 9. The final set of these sentences created can be seen in Table 3.4.

Table 3.4: Complex sentences created for the impaired speakers

Sentences
El *toro* en el *árbol* bajo la *luna*
El *ratón* roe el *plátano* sobre la *silla*
La *bruja* da *pan* a la *cabra*
El *niño* se *ducha* con su *gorro*
Dejaba el *periódico* en la *puerta* de *casa*
El *indio* tocaba la *campana* del *tren*
La *tortuga* abría la *boca* bajo el *sol*
Perdí la *llave* de la *jaula* del *preso*
La *mariposa* en la *palmera* de la *playa*
Tocar el *tambor* con *pala* y *cuchara*

The difficulties for the production of this complicated sentences were high, as speakers had problems to utter them, not only due to their phonological difficulties, but for their cognitive disorders, which made complicated for them to comprehend fully the sentences; which provided extra problems in the form of hesitations or doubts during the recordings. The total number of recorded utterances was, then, 40 for a total time of 2 minutes and 9 seconds of speech including silence. The mean SNR per utterance was 33.21 dBs with an standard deviation of 5.34 dBs.

3.5 Reference Speech Corpus

Speech technologies usually require of well-matched data to achieve the best performance. Speaker adaptation is the best solution, but it is not always available and sometimes not desirable. Even in those cases, it is interesting to be able to model properly the task and domain in which the system will work; this is, the vocabulary used, the acoustic environment, the dialect or some other characteristics of the speakers that the system will face.

For the corpus of disordered speech presented in this Chapter, there were several elements that create a mismatch with the traditional ASR corpora used in Spanish like Albayzín [Moreno et al., 1993]. Main one was, obviously, the disorders that the speakers in the Alborada-I3A corpus suffered in their speech and language. But, difference of age was also a main element that produced a mismatch between the corpora. In this case, this age mismatch could cover effects of the speech impairments because children' voices as the one in the Alborada-I3A corpus were always expected to obtain worse performances than adult voices in ASR. Hence, it was shown the need of counting with a reference speech corpus from unimpaired children.

This children speech was donated by students from three different institutions in Zaragoza: The Colegio de Educación Infantil y Primaria (School for Primary and Infant Education) (CEIP) "Río Ebro", the Instituto de Educación Secundaria (School for Secondary Education) (IES) "Tiempos Modernos" and the IES "Felix de Azara", with the consent of the parents and educators of the children.

Recording environment was the same that the one for the impaired speakers shown in Section 3.2; this is, speakers were recorded in the facilities of the schools were they attended their classes, and each speaker was asked to utter only one session of the RFI isolated words like in Section 3.4.1. Balance in age and gender was the target in the same range of ages that the impaired speakers; and finally, the distribution was as shown in Table 3.5.

Table 3.5: Speakers in the reference speech corpus

Age	Males	Females	Age	Males	Females
10 years old	15	16	11 years old	15	16
12 years old	15	15	13 years old	15	23
14 years old	11	21	15 years old	11	11
16 years old	15	9	17 years old	14	10

With this distribution, the total number of unimpaired speakers in the reference subcorpus was 232 with 13224 utterances (one session per speaker), and a total signal time of 8 hours, 50 minutes and 15 seconds of speech including silence. The average SNR per utterance was 33.27 dBs with a standard deviation of 6.0 dBs; similar values to the SNR values of the impaired speech subcorpus in Section 3.4, as it was expected because the acquisition scenario was similar.

3.6 Human Labeling of the Oral Disorders Corpus

The second of the extra tasks that was made to complete the corpus was to manually label the mispronunciations in the corpus. This task was required to allow the evaluation of the performance of algorithms for PV or to study the impact of mispronunciations in the performance of different proposed ASR systems.

An important decision prior to the labeling was to decide the possible labels that were going to be the outcome of the labeling. An accurate full quality transcription would be the most precise labeling possible; in this labeling, the output would be the whole real transcription of the words uttered by all the speakers with a mark measuring the speech quality of the pronunciations of that phoneme. Phoneme and word boundaries would be also provided by the labelers; giving, this way, a manual accurate segmentation. Despite the accurateness of this type of labeling, a major drawback would be that the consistence among different labelers might be very low as it would rely on subjective opinions of each one of them. This lack of agreement would give very little statistical significance to the results of the labeling, limiting the usefulness of it.

Furthermore, with this kind of labeling, more information would be obtained than really necessary; as it was not intended to make a speech quality assessment task. The main reason for this is that the speakers are children with moderate to severe cognitive disorders; in those cases, the final objective of speech therapy in special education is to provide them with communicative skills, and communication is not on perfect quality but the perceptual in discrimination of different phonemes. Hence, real interest in this task is to distinguish whether human listeners could accept words according to the canonic transcription or not.

For this reason, the ultimate decision was that the label that the human experts were to assign to each phoneme could only be 0 (phoneme was deleted), 1 (phoneme was substituted by another phoneme or by an unintelligible pronunciation) or 2 (phoneme correctly pronounced as the canonic phoneme independently of the quality of the pronunciation). When the experts marked a 1 for a given phoneme, they were not asked initially to provide which phoneme they understood to have been pronounced instead of the canonical phoneme to avoid delays in the labeling and inconsistencies. A phoneme was considered as mispronounced when it was considered as substituted (1) or deleted (0).

The labeling of the corpus had to be reliable as well as significant. For this purpose, a group of 14 labelers were chosen from disciplines like speech technology or phonetics. Each one of the 56 sessions (14 speakers and 4 sessions per speaker) was handed to 3 different labelers out of the group of 14 experts that evaluated separately all the words. A same labeler never had to face more than two sessions of a same speaker, to avoid two effects: that the labeler could change the results as the labeler got adapted to the speaker's speech; and, furthermore, that a same expert had too much influence in the labeling results of an speaker.

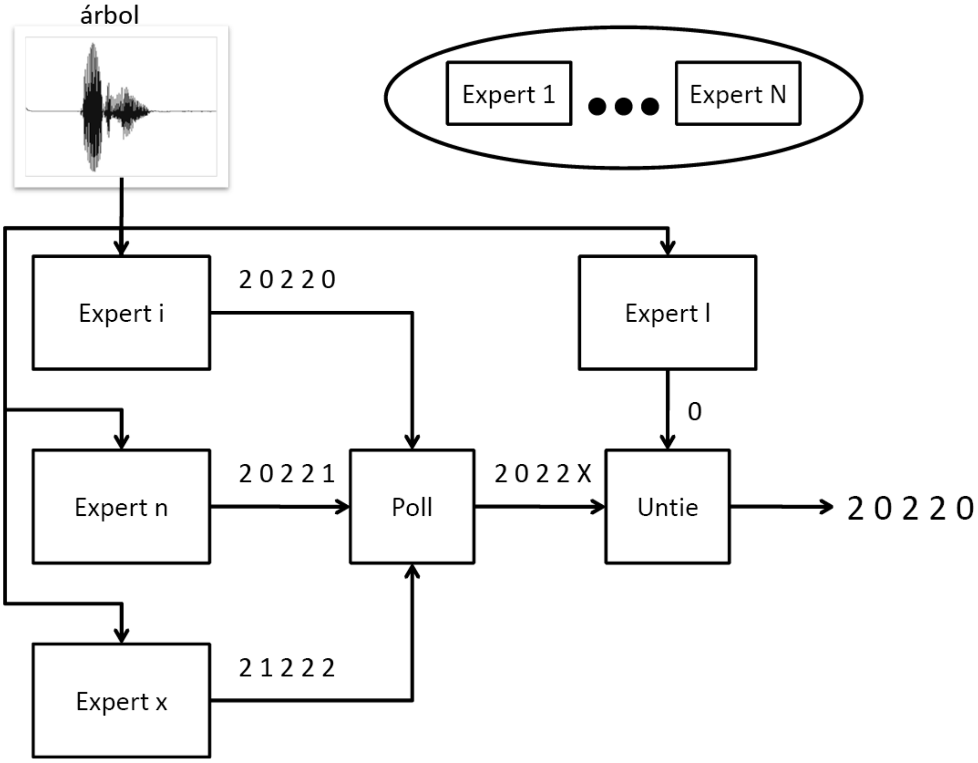


Figure 3.5: Illustration on the labeling process

The orthographic transcription of each word was presented to the labeler and the expert could listen to the speaker's utterance as many times as needed prior to give the label (0, 1 or 2) over all the phonemes in the canonic phonetic transcription of the word. Finally, with a polling system, the most voted label by the three labelers was chosen as the final label of the phoneme. In the case that a phoneme receive three different values by the three different experts, another labeler was requested to provide a label to that phoneme and untie the result.

Example 3.6.1 An example illustrating the labeling process is shown on Figure 3.5. In this case, the word to label is árbol (*tree*), whose transcription is in 5 phonemes (/a/, /ɾ/, /b/, /o/ and /l/). In the example, Expert i labels phonemes /ɾ/ and /l/ as deletions, while the other are corrects; Expert n labels /ɾ/ as deleted and /l/ as substituted; and, finally, Expert x only labels /ɾ/ as substituted. The polling system assigns phonemes /a/, /b/ and /o/ as correct as all three labelers agree; and also assigns phoneme /ɾ/ as deleted because 2 out of 3 labelers agree. Finally, phoneme /l/ cannot be assign a final label, as there is a tie between all three labelers. Expert l is called then and asked to label that phoneme; his label is 'deletion' and this is the final label assigned to phoneme /l/.

The outcome of the labeling in terms of correct phonemes, substituted phonemes and deleted phonemes by speaker is shown on Table 3.6. The final average values for the total corpus (the 4 sessions of the 14 speakers) was 82.39% of correct phonemes, 10.31% of substituted phonemes and 7.30% of deleted phonemes. A measure of the average rate of mispronunciation for all the speakers was 17.61%, considering as mispronounced phonemes the sum of substituted phonemes and deleted phonemes.

Table 3.6: Results of the labeling process
Percentage of correct, substituted and deleted phonemes in the corpus

Speaker	Correct	Substituted	Deleted	Speaker	Correct	Substituted	Deleted
*Spk*01	98.88%	0.94%	0.17%	*Spk*02	78.42%	12.41%	9.16%
*Spk*03	94.78%	4.54%	0.68%	*Spk*04	96.83%	2.05%	1.11%
*Spk*05	56.51%	26.11%	17.38%	*Spk*06	99.32%	0.51%	0.17%
*Spk*07	87.07%	7.36%	5.57%	*Spk*08	69.18%	17.72%	13.10%
*Spk*09	91.78%	5.31%	2.91%	*Spk*10	78.51%	13.10%	8.39%
*Spk*11	93.24%	5.15%	2.05%	*Spk*12	74.32%	13.96%	11.73%
*Spk*13	43.58%	30.48%	25.94%	*Spk*14	91.01%	5.14%	3.85%

To measure the consistency in which different human experts can identify the phonetic accuracy of the utterances in the corpus the pairwise interlabeler agreement was calculated. This measure was obtained by the comparison of every two possible pairs of labels assigned by the experts, which was obtained comparing every two possible pairs of labels for a given phoneme. The total number of phonemes to label was 16 352, which supposed 49 056 comparisons between experts' labels pairwise. Labelers agreed in 42 095 cases, which marked a 85.81% of agreement. This meant that any automatic system that tried to imitate the experts would achieve a reliable performance (similar to the humans) when reaching 85% of accuracy. If only two levels were to be considered: correct (marked as 2) and mispronounced (marked as 1 or 0) phonemes; the interlabeler agreement raised to 89,65%.

Example 3.6.2 Following the Example 3.6.1 and Figure 3.5, the agreement rate could be easily calculated for that case. There are five phonemes to evaluate and there are three possible pairwise interlabeler comparisons (Expert i with Expert n, Expert i with Expert x and Expert n with Expert x), this makes 15 total comparisons. Situations in which the experts disagree are 5 (Expert i with Expert x in phoneme /ɾ/, Expert n with Expert x in phoneme /ɾ/, Expert i with Expert n in phoneme /l/, Expert i with Expert x in phoneme /l/ and Expert n with Expert x in phoneme /l/). Hence, interlabeler agreement is 10 out of 15 situations (66.67%).

The last measure of the labeling consistency was the rate of phonemes in which a fourth expert was required; this only happened in a 1.38% of all the phonemes in the corpus (226 out of 16,352 phonemes). It is to remind that this was a very unexpected situation, as it meant that labelers were giving extremely contradictory labels to the same phoneme, although it might happen in extreme cases where the pronunciation was difficult to assess for the labelers.

3.7 Acknowledgments

An special acknowledgment has to be given in this Chapter for all the people that has gratuitously collaborated in the acquisition of the corpora explained in the Chapter.

First, to the staff, educators and pupils of the CPEE "Alborada", especially to José Manuel Marcos Rodrigo, Cesar Canalís, Pedro Pegero and Beatriz Martínez for their support and for allowing the recordings of the 14 students that are present in the disordered speech corpus.

Extra thanks for the reference corpus recordings to Nuria Lozano, from the CEIP "Rio Ebro" for providing the children in the range of 10 to 12 years old and to Pilar Burillo from the IES "Tiempos Modernos" and Joaquín Lostal from the IES "Félix de Azara" for providing the speakers of 12 to 18 years old in this reference corpus.

Chapter 4

Experimental Framework and Baseline Results

> The box.
> You opened it.
> We came.
> Now you must come with us, taste our pleasures.
>
> -Clive Barker, *The Hellbound Heart*

Characterizing the task in which an ASR system is aimed to work is strongly required prior to start the study of further improvement methods. This Chapter aims to set the baseline ASR performance on the disordered speech corpus introduced in the previous Chapter. The objectives are to detect sources of variability in this speech and evaluate their effect on the ASR recognition rates, while considering different possibilities in HMM topologies.

The Chapter is organized as follows: In Section 4.1, the presentation of the different HMM possible topologies and the feature extraction method that was used in the experimental framework will be made. Sections 4.2 and 4.3 will present the results in ASR in the impaired children' speech corpus in two different approaches: TI and TD, comparing these results to the results obtained in the same tasks by the unimpaired age-matched speakers. A small review on the results in ASR with connected disordered speech will be given in Section 4.4, as well as the baseline results in APD in Section 4.5. Finally, a discussion on the influence of the labeled mispronunciations in the disordered speech corpus on the presented results will be given in Section 4.6.

4.1 Description of the Baseline ASR System

A correct selection of the characteristics in the ASR system to be used in a given task and domain is necessary to achieve the best possible performance. The aim of this thesis is to personalize a traditional ASR system towards disordered speech, so all the experiments in the thesis were run on a state-of-the-art system, throughly used and evaluated in many tasks with adult normal and healthy speech like digits [Buera et al., 2007, Miguel et al., 2008], command control [García et al., 2008] and broadcast news subtitling [Ortega et al., 2009], with results comparable to those of well-known libraries like Hidden Markov Model Toolkit (HTK) [Young et al., 2006] or Sphinx [Lee, 1989]. No modifications on the Viterbi search or in the HMM left-to-right topology were made from the original framework, to make possible the use of the system by both unimpaired and impaired speakers indistinctly.

4.1.1 Selection of the HMM topology

HMMs have been used for a long time in ASR to model and characterize human speech [Jelinek, 1998]. This statistical modeling fits the pseudo-stationary properties of the speech signal and allows the decoding of the uttered message via forward-backward algorithms in a Viterbi search [Huang et al., 2001]. Traditional HMM topologies follow a left-to-right structure like the one shown in Figure 4.1. In this topology, the model follows the speech variations in time from one state to the next, where the time of latency in each state is defined by the probability of permanency. Each state in the HMM is defined by a probability density function (pdf) (usually a GMM with a certain number of Gaussian distributions) that sets the probability with which a given speech frame has been generated by that state.

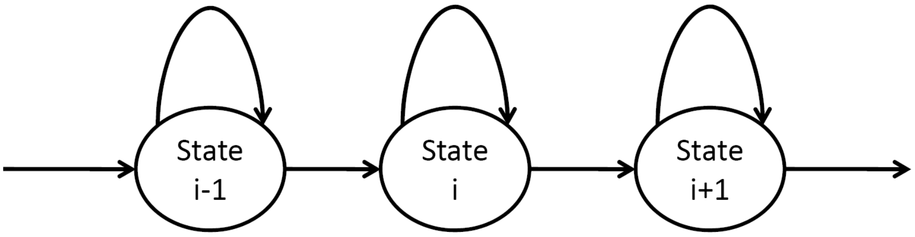


Figure 4.1: Left-to-right HMM topology

To define the HMM topology used in an ASR experimental framework, it is necessary to provide the acoustic units that are modeled with each HMM, the number of states per HMM and the number of Gaussian mixtures per state. An acoustic unit is the segment of speech that is modeled with a different HMM and hence defines the smaller units in which the recognition can be divided.

Three types of acoustic units were considered initially for the experimental framework in this thesis: Word units, phoneme units and sub-phone (context dependent) units. Each one of them has its own advantages and disadvantages in different tasks and domains that will be seen along this Chapter and thesis.

Word units model acoustically a whole word and, hence, a new unit is required each time a new word is included in the vocabulary. They have a great specificity, as they are trained to model one and only word. Furthermore, the training of a new word unit can not be assured as the word might not appear in the training database and requires the whole re-training over the database. The number of states per word HMM is variable as the length of the different words might highly vary. The number of word models included in the works of this thesis was 57, one for each word in the RFI.

Phoneme units are trained to model a given phoneme or allophone, independently of its left and right phonetic contexts. These units are very generalizable, as every possible word in the vocabulary can be described as a concatenation of a small set of units (Spanish has 24 phonemes and about 50 allophones). However, they are very little specific, because it is well known the high acoustical influence of the context in the production of the different phonemes (part of this can be avoided with the use of an extended set with all the allophone units). The final number of phoneme units in the experiments in this thesis was 25, covering 23 of the 24 phonemes of Spanish (phonemes /ʝ̞/ and /ʎ/ are gathered in the lateral sound [ʎ] due to the common 'yeísmo' explained in the Appendix A) and the two glides [j] and [w], allophones of the vowels /i/ and /u/ respectively.

The reduction of /ʝ̞/ to /ʎ/ was justified by the fact that all speakers in the corpus came from the North-Eastern region of Spain, where yeísmo is fully expanded [Frago-Gracia, 1978], which made the speakers of the corpus unable to separate both phonemes in their speech. Furthermore, all labelers came also from the same geographical distribution, which limited their ability to distinguish the differences between /ʝ̝/ and /ʎ/ as a mispronunciation if that would have happened in the speakers' utterances.

On the other hand, separation between glides and their corresponding vowels has shown a significant improvement in ASR performance in Spanish because their role on syllable construction differs extremely when the vowel phoneme is nucleus or glide. Moreover, this separation was a necessary requirement for the analysis carried out at the acoustic and lexical levels of the quality in the speakers speech. Glides will be separated from vowels in the acoustic analysis in Chapter 5 and glides in diphthongs will be shown to have a major influence in the lexical mistakes of the speakers in Chapter 6. However, in any case, labelers were never required to distinguish both sounds (vowel from glide) in their annotation, so for PV purposes no distinction was made amongst then.

Finally, *sub-phone* units split phoneme units into several smaller units that consider the context of the given unit. Three type of units are considered then: left context units, which model the beginning part of a phoneme with the current left context; center context units, that model the context-independent part of the phoneme in the middle of the phoneme; and the right context units, modeling the ending part of the phoneme towards the phoneme to the right. Sub-phone units are highly specific, as they consider an specific context of the unit. They are also generalizable, because every new word in the vocabulary can also be created from these units. However, sub-phone units are difficult to train as the number of possible units is in the order of the number of phonemes squared, creating possible situations of sub-training if there are not enough realizations of the unit in the training set. The final number of context-dependent sub-phone units trained for the works in the thesis was 744, although only 309 were actually used in the recognition experiments due to the small size of the vocabulary.

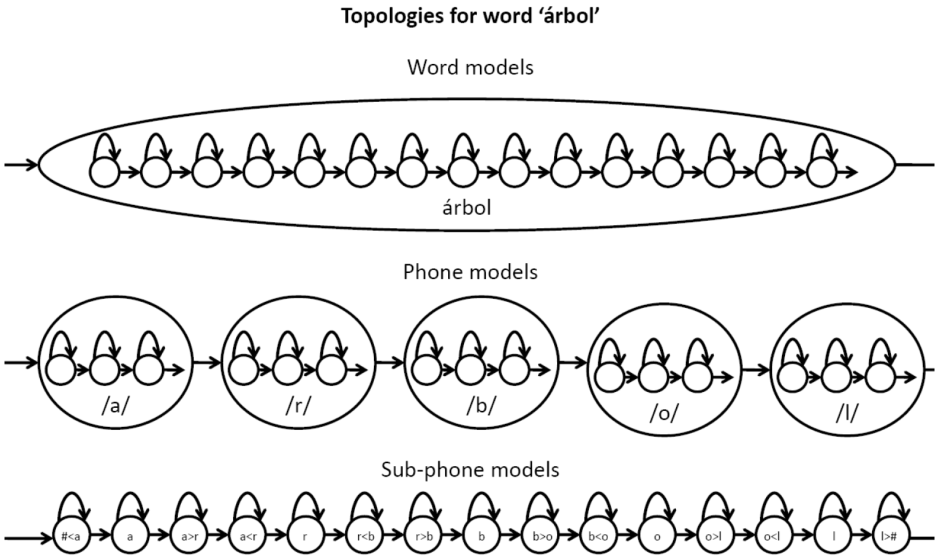


Figure 4.2: Examples of the different HMM topologies

Other possible units like syllable units or triphone units, widely used in ASR, were not considered for this job, as they shared similar properties with the proposed units. While syllable units model long contexts like word units, triphones consider the left and right contexts like the sub-phone units.

Example 4.1.1 An example illustrating the possible topologies is shown on Figure 4.2. In this case, the different possibilities to model the word árbol (*tree*) are shown. With word units, an only model is created with the whole word and 15 states for this model. With phone units, the 5 models for /a/, /ɾ/, /b/, /o/ and /l/ are used to construct the word; each phoneme unit with 3 states. Finally, 15 sub-phone units are required to construct the same word, with only 1 state per model.

Training of the HMMs used is usually made via large transcribed databases, where the number of appearances of each unit is high enough to provide enough examples in all its relevant possible contexts. The training is based on the solution of an Expectation-Maximization (EM) problem that finally achieves a Maximum Likelihood (ML) solution [Dempster et al., 1977].

The initial models for this work were trained with data from three major speech corpora in Spanish containing adult clean healthy speech: Albayzín [Moreno et al., 1993], Spanish SpeechDat-Car [Moreno et al., 2000] and Domolab [Justo et al., 2008]. With 44,018 sentences from several speakers in different tasks, these corpora were consistent enough for the training of baseline acoustic models in different tasks like ASR in clean [Miguel et al., 2008] and noisy conditions [Buera et al., 2007] or speaker verification. With these corpora, the phone and sub-phone models were trained. Baseline word models were constructed by the concatenation of sub-phone units, because they could not be consistently trained for all the words in the RFI vocabulary from these databases due to the lack of training data for some of them.

4.1.2 Feature extraction method

The speech signal processing for ASR was based in the extraction of a group of features that describe the articulatory properties of each frame of speech in the best and most compact possible way. A Mel Frequency Cepstral Coefficients (MFCC)-like processing was applied to the input speech signal following the flow diagram in Figure 4.3.

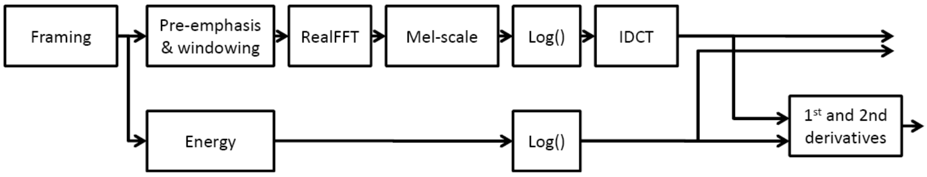


Figure 4.3: Feature extraction method

Each input speech signal was framed into 25msec. frames with an overlap of 15msec. A pre-emphasis filter was used to discard the possible continuous component of the speech signal and the very low frequencies. Posteriorly, a Hamming window was applied to each frame and zero-padded to 512 samples per frame to apply the Fast Fourier Transform (FFT) to the frame. The real part of the FFT was converted to a 24-bin Mel scale and converted to logarithm scale. With the Inverse Discrete Cosine Transform (IDCT), the Mel-scale parameters were transformed to the cepstral domain, and the first 12 cepstral coefficient, discarding $c0$, were used ($c1$, ..., $c12$). The logarithm of the energy of the input frame was calculated and added to the cepstral coefficients instead of $c0$.

Posteriorly, the first and second discrete derivatives of the obtained 13 parameters per frame were calculated and the final 39 MFCC features were fed to the training and recognition systems.

4.2 Task Independent ASR Results

With the MFCC-based feature extraction method presented and the three HMM topologies proposed for study, the first ASR experiments had to be oriented to determine the baseline of the recognition system in the oral disorders corpus. Initially, the reference speakers provided the knowledge on how accurate the models were for children' speech.

4.2.1 Reference speakers' results

The 232 reference speakers of children and young speech were fed to the ASR system defined previously. The initial Speaker Independent (SI)-TI models trained as commented in the previous Section were used for recognition, with the results shown in Table 4.1. All the ASR results in this thesis are provided in terms of WER, but in the presence of isolated words as it is the case, only the substitution of words was possible and, hence, the standard WER equation was simplified to Equation 4.1.

$$WER = \frac{Substitutions}{Words} \tag{4.1}$$

Table 4.1: Baseline ASR results for the reference speakers with the TI models

WER results for the unimpaired speakers

Word models	Phoneme models	Sub-phone models
3.99%	9.91%	3.99%

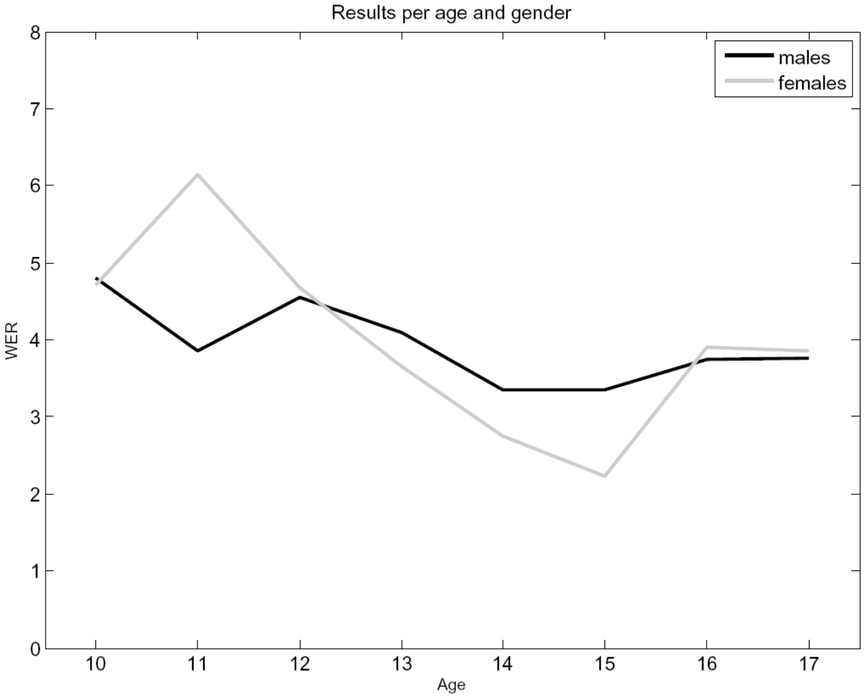


Figure 4.4: ASR results of children' speech in terms of age and gender

As it could be expected, word and sub-phone models outperformed the phoneme models due to the better modeling that they did of the context and coarticulation of phonemes. The results of word and sub-phone models were exactly the same because, as it was explained in previous Section, word models were created from the concatenation of sub-phone units.

Speech from children and young adults is known to be very variant depending on the age of the speaker. Different speakers in these ages present different vocal tract and vocal chords features (lengthening of the vocal tract and lowering of the fundamental frequency with growth) which may produce very different performance of the ASR systems. To study the impact of these effects, the results obtained in the ASR system with the reference corpus were separated in terms of age and gender to know how this variability affected them. Figure 4.4 shows these results per age and gender for the sub-phone models. Variations in the WER through age and gender marked a minimum decrease for the older speakers (especially in female speakers). Anyways, the variations were not so significant to indicate a trend with age or gender, probably due to the good modeling of the different units, the short size of the vocabulary and the controlled environment of the recordings. A task more oriented to spontaneous speech would suffer more of this age effects in the speech.

4.2.2 Impaired speakers' results

The same experimental framework was posteriorly evaluated over the 14 impaired speakers. The special characteristics of these speakers exposed out the need of providing full results separately per speaker; the results for these speakers are provided in Table 4.2. Again, results for word and sub-phone models were the same as the initial TI word models were created from the concatenation of sub-phone units. The average (*AVG*) results for all speakers (36.69%, 40.86% and 36.69%) showed a significant loss of performance in comparison to the unimpaired speakers (3.99%, 9.91% and 3.99%).

Table 4.2: Baseline ASR results for the impaired speakers with TI models

WER results for the impaired speakers

Speaker	word	phone	sub-phone	Speaker	word	phone	sub-phone
*Spk*01	16.23%	19.30%	16.23%	*Spk*02	39.47%	40.79%	39.47%
*Spk*03	21.49%	32.02%	21.49%	*Spk*04	13.60%	18.42%	13.60%
*Spk*05	63.60%	63.16%	63.60%	*Spk*06	6.58%	16.67%	6.58%
*Spk*07	33.33%	37.72%	33.33%	*Spk*08	42.54%	46.05%	42.54%
*Spk*09	36.40%	38.16%	36.40%	*Spk*10	43.86%	50.44%	43.86%
*Spk*11	13.60%	21.93%	13.60%	*Spk*12	69.30%	72.37%	69.30%
*Spk*13	82.46%	79.39%	82.46%	*Spk*14	31.14%	35.53%	31.14%
AVG	36.69%	40.86%	36.69%				

The high variability in the results pointed out the very different performance obtained by all the speakers, which was due to the very different personal situation of each speaker. Results ranged from the 6.58% and 16.67% of *Spk*06 to the 82.46% and 79.39% of *Spk*13 in WER. It was interesting to see how phoneme units did not perform so badly in comparison to the highly specific sub-phone units in the impaired speakers (40.86% vs. 36.69%) than in the unimpaired speakers (9.91% vs. 3.99%). Main explanation for this might arise from the loss of accurate articulatory properties in the disordered speech, which produced that the high accuracy of the sub-phone units was not so relevant in the speech from speakers whose articulation is in many cases different from the baseform pronunciation.

4.3 Task Dependent ASR Results

Task and domain adaptation is an easy and direct way to improve the performance of all ASR systems. If the vocabulary expected to be uttered by the speakers (task) is fixed and the acoustic context and characterictics of the speakers (domain) are also fixed; it is possible to train or adapt the acoustic models to this specific task and domain situation.

In the situation proposed in this thesis, the task is fixed to the 57 words in the RFI and the domain are the special properties of speech from children and young adults recorded in a clean acoustic environment. Hence, task and domain adaptation could initially provide a very interesting improvement in the ASR results.

This adaptation was carried out by means of the Maximum A Posteriori (MAP) algorithm [Gauvain and Lee, 1994]. MAP can provide a very good adaptation (similar to ML when sufficient data is available for adaptation). A small review on adaptation algorithms like MAP or Maximum Likelihood Linear Regression (MLLR) is provided on Appendix B, to provide the reader a view on the final implementation of both algorithms.

4.3.1 Reference speakers' results

For studying the improvement that TD models could provide in the recognition of the reference speech, two different tests had to be carried out to assure independence between the speakers used for training and for testing. These two tests were defined by the creation of two subsets in the reference speech corpus as seen in Table 4.3 that kept balance in age and gender as the original full reference corpus.

Table 4.3: Speakers in the subsets of the reference speech corpus

	Subset A		Subset B	
Age	Males	Females	Males	Females
10 years old	8	8	7	8
11 years old	7	8	8	8
12 years old	8	8	7	7
13 years old	7	11	8	12
14 years old	6	11	5	10
15 years old	5	5	6	6
16 years old	8	5	7	4
17 years old	6	5	8	5
Total	55	61	56	60

With this separation, two TD models were trained with the 116 speakers in each subset via the MAP algorithm. Each one of these two models was used to recognize the other subset. The final WER results in Table 4.4 were obtained as the average of the results obtained in both experiments.

Table 4.4: Baseline ASR results for the unimpaired speakers with the TD models

WER results for the unimpaired speakers		
Word models	Phoneme models	Sub-phone models
2.04%	2.77%	2.11%

The results showed an improvement which lowered the WER to 2-3%, depending on the acoustic units. The impressive improvement achieved by TD phoneme models (from 9 to 3%) could be explained by the small size of the task vocabulary. With this small vocabulary, the number of different contexts in which the phonemes appeared got extremely reduced and could be better modeled by context independent models like phonemes.

4.3.2 Impaired speakers' results

The results with the TD models trained from the whole reference speech are in Table 4.5 for the three units (word, phoneme and sub-phone). The improvement achieved by the task and domain adaptation reached in some speakers very impressive values, up to 75% of reduction in WER for *Spk*04; while in the extremely impaired speakers like *Spk*13, it only implied a 5% improvement. The average WER was lowered down up to 25.85%, 31.96% and 28.20% with word, phone and sub-phone units respectively, marking the baseline for all the experiments performed in the thesis to try to obtain an improved recognition to these speakers.

Table 4.5: Baseline ASR results for the impaired speakers with TD models

WER results for the impaired speakers

Speaker	word	phone	sub-phone	Speaker	word	phone	sub-phone
*Spk*01	6.14%	11.40%	10.09%	*Spk*02	16.67%	22.81%	20.18%
*Spk*03	5.26%	14.91%	5.70%	*Spk*04	3.07%	4.39%	3.51%
*Spk*05	58.77%	60.09%	56.14%	*Spk*06	4.39%	7.46%	3.07%
*Spk*07	22.37%	32.02%	25.44%	*Spk*08	39.91%	46.49%	43.86%
*Spk*09	22.81%	25.00%	22.81%	*Spk*10	22.37%	38.16%	32.46%
*Spk*11	7.46%	13.60%	9.21%	*Spk*12	57.02%	70.61%	63.60%
*Spk*13	77.63%	77.19%	80.26%	*Spk*14	17.98%	23.25%	18.42%
AVG	25.85%	31.96%	28.20%				

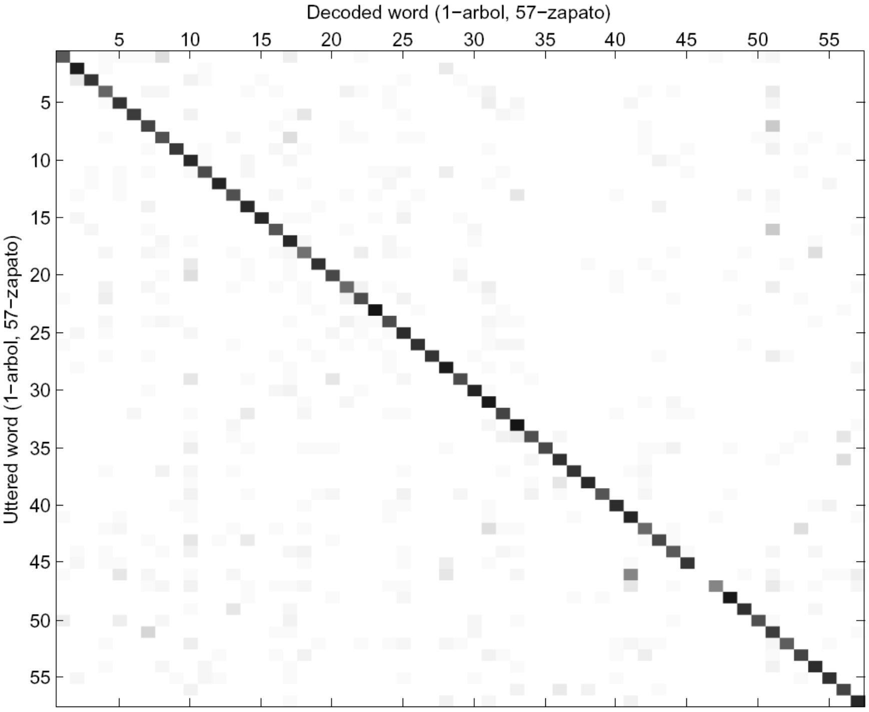


Figure 4.5: Confusion matrix for the ASR of the 14 impaired children

More information on the way the recognition system was producing the misrecognitions could be observed in the confusion matrix for the 14 speakers in the experiment with word units, shown in Figure 4.5. This matrix showed up how the mistakes were being produced quite uniformly over the matrix, with no specific pattern for the 14 speakers, apart from a noticeable poor performance for the recognition of the word "ratón". Anyways, Appendix C provides more information with the confusion matrices of all the speakers in this recognition experiment. These confusion matrices were consistent with the WER values showed by each speaker, and some of them showed very interesting phenomena of reduction of words.

Word units showed the best performance in the TD system, showing that with a short vocabulary like the RFI they could achieve a good modeling of the coarticulation effects. Unfortunately, their inability to adapt to new lexicon entries in the vocabulary (either they are new words or new transcriptions of old words) will be seen in further Chapters in the thesis as an obstacle for the further research in personalized systems where vocabulary could change dinamically or where new lexicon variants could be introduced in the vocabulary. This drawback did not exist for the other units, sub-phone and phone units, which, especially in the case of sub-phone units, did not lose much performance compared to the outstanding word models.

4.3.3 Domain adaptation to disordered speech

Another possible framework for task and domain adaptation was to adapt the models to the impaired speakers domain within the same 57 RFI task. This domain was characterized by the set of speakers in the disordered speech corpus, who shared similar conditions in age and were all affected by different speech disorders. To evaluate this approach, 14 different models were trained using the baseline TI models as seed. Each model used speech from the 4 sessions of 13 of the impaired speakers and was evaluated in recognition over the 4 sessions of the remaining speaker.

The results in Table 4.6 did not show a significant difference with the TD models trained from the reference unimpaired speakers. Hence, it could be seen that there was not further effect of adaptation to the disorders of the speakers and that the overall effect was an adaptation to the characteristics of the age of the speakers. This was probably due to the very different nature of the disorders of each speaker, which made unable to create an acoustic model that could provide effective adaptation for all the range of possible disorders existing in the corpus.

Table 4.6: Baseline ASR results for the impaired speakers with disordered TD models

WER results for the impaired speakers							
Speaker	word	phone	sub-phone	Speaker	word	phone	sub-phone
*Spk*01	6.58%	7.02%	7.46%	*Spk*02	21.05%	24.12%	25.88%
*Spk*03	13.60%	24.56%	17.54%	*Spk*04	4.83%	8.33%	5.26%
*Spk*05	53.51%	56.14%	50.44%	*Spk*06	8.33%	17.54%	11.40%
*Spk*07	21.05%	27.63%	23.68%	*Spk*08	41.23%	47.37%	47.37%
*Spk*09	23.25%	21.93%	20.61%	*Spk*10	23.68%	25.88%	26.75%
*Spk*11	7.89%	20.61%	10.96%	*Spk*12	62.72%	74.56%	64.04%
*Spk*13	74.12%	75.44%	75.88%	*Spk*14	15.79%	14.91%	13.16%
AVG	26.97%	31.86%	28.60%				

As the whole group of impaired speakers did not show the ability of being characterized by itself, the need of personalization appeared stronger than ever and provided further motivation for the overall objective in the thesis of building fully personalized systems in future Chapters.

4.4 Results in Connected Speech

This Section wants to bring a small introduction to the influence of language impairments in disordered speech. The comparison of the isolated word results and the connected speech results might indicate the possible extra difficulties that impaired speakers face when uttering full sentences. Unfortunately, the results in this Section could not be considered significant for obtaining conclusions because the amount of data in isolated words was much bigger than the amount of data in connected speech. The special characteristics of these speakers did not allow to acquire more of this data in the time of the recordings as it was seen in Chapter 3.

The experimental framework in connected speech was divided in simple meaningless sentences and complex meaningful sentences as it was made during the recordings explained in Section 3.4.

4.4.1 Simple meaningless sentences

The simple meaningless sentences defined in the corpus followed a direct construction from the words in the RFI. For the recognition experiments of these sentences a new grammar had to be created to expand the isolated-word grammar used in the previous experiments. This new grammar included the rules that defined the meaningless sentences and followed the flow diagram in Figure 4.6. The connecting words ('el', 'la' and 'y') were forced into the grammar for recognition and not considered for computing the WER. With this approach, the WER obtained in the sentences could be compared to the WER of isolated words where only the RFI words were recognized.

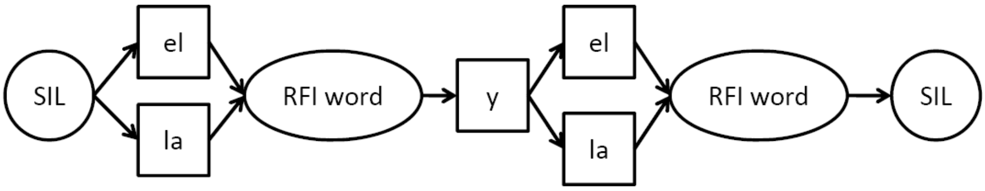


Figure 4.6: Grammar for the recognition of the simple meaningless sentences

A drawback in the change of task from isolated words to connected speech was the inability to use word models. As it was explained at the beginning of this Chapter in Section 4.1, word acoustic models cannot be generalized to introduce new words in the vocabulary like the connectors that were used in this task. Hence, only phoneme and sub-phone models could be used by introducing the new words in the vocabulary with their phonetic expansions. Although, the new word units could have been created directly from sub-phone units like it was made with the baseline units, it was decided to maintain in all cases the original units to reflect a more realistic case, in which new word units cannot be created directly from previous word units.

Table 4.7: Baseline ASR results for the impaired speakers

Model	*Spk*01 WER	*Spk*04 WER	*Spk*06 WER	*Spk*11 WER
Phoneme TI	16.37%	26.55%	7.96%	25.22%
Sub-phone TI	15.04%	18.58%	5.31%	14.16%
Phoneme TD	13.72%	24.78%	11.06%	24.34%
Sub-phone TD	14.16%	18.14%	6.19%	15.04%

The results in the simple meaningless sentences can be observed in Table 4.7. Comparison between the results of TI models in isolated words and connected speech showed similar results, except for *Spk*04, who suffered a major increase in WER for phoneme and sub-phone models. This result in this speaker might indicate that the degree of her language disorders was seriously degrading the quality in her speech in the change from isolated to connected speech. However, there was no sufficient data in this new task to make any ultimate conclusion about this. Any conclusions about these results had to be kept, hence, as pure suppositions as it was necessary to remark again the insufficient data for this task.

Regarding the TD models, there was a major loss of performance in connected speech compared to isolated words, resulting in a lack of improvement between TI and TD models. However, it was noticed that the condition of task dependence was not strictly accomplished this time, because these models were trained with isolated words from young unimpaired children and did not cover the special coarticulation effects appearing in a task of connected speech.

4.4.2 Complex meaningful sentences

For the recognition of the complex sentences, a new grammar was created adapted to this task following the diagram in Figure 4.7. These sentences were richer that the simple sentences in terms of the words that appeared in them; and no ad-hoc grammar was created this time and no Statistical Language Model (SLM) was created as no corpora was available that could model properly the special types of sentences created with the words in the RFI. The proposed grammar allowed for computing the recognition mistakes of the system in the 3 RFI words included in each sentence while the phoneme network forced the recognition of the phonemes of the rest of the words.

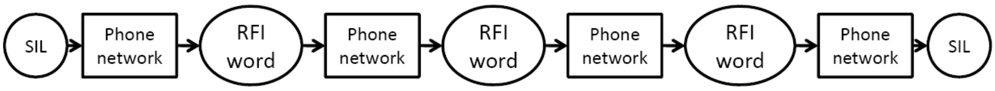


Figure 4.7: Grammar for the recognition of the complex meaningful sentences

The results in this task are presented in Table 4.8 showing a serious loss of performance when compared to the isolated words or the simple meaningless sentences. However, these results could not be generalized due to the small amount of data for the experiments (only 30 words for recognition from each speaker). These poor results were also influenced by the fact that no specific recognition system was prepared for the connected speech task and the recognition was limited to the RFI words with the Out-Of-Vocabulary (OOV) words just modeled as phone sequences.

Table 4.8: Baseline ASR results for the impaired speakers

Model	*Spk*04 WER	*Spk*06 WER	*Spk*11 WER
Phoneme TI	63.3%	56.7%	46.7%
Sub-phone TI	60.0%	36.7%	40.0%
Phoneme TD	53.3%	33.3%	43.3%
Sub-phone TD	43.3%	36.7%	46.7%

4.5 Acoustic Phonetic Decoding (APD)

APD aims to decode the most probable sequence of phonemes uttered by the speaker [Lee and Hon, 1989] in a utterance. An APD system can be seen as an ASR system where the words in the vocabulary are the phonemes of the language. The uses of APD can be many like the decode of OOV words like words in a foreign language or proper names.

In the case of disordered speech, it can be useful to compare the decoded phoneme sequence with the canonical transcription and determine if possible dissimilarities are appearing because of possible mispronunciations due to the speakers' disorders. These differences obtained by the output of the APD framework could be closer to possible lexical variants introduced by these special speakers as it will be seen later.

The outcome of the APD was measured in terms of PER. The PER considers all possible mistakes in recognition: substitutions, insertions and deletions, following the tradition equation for computer recognition mistakes in Equation 4.2.

$$PER = \frac{Insertions + Substitutions + Deletions}{Phonemes} \tag{4.2}$$

4.5.1 Phonotactic language modeling for APD

The phonotactic language model in an APD system models the way in which phonemes get together in a language to create syllables and words in a similar way in which a grammar in ASR models the way in which words get together in the target language or task to create sentences.

Two strategies for language modeling in APD were considered: First, a data-driven stochastic grammar with bigrams and trigrams was trained from 700,000 sentences of the Spanish subset of the Europarl corpus [Koehn, 2005], with more than 18 million words and more than 94 million phonemes. A total of 628 bigrams and 9110 trigrams were created for the grammar with phone units and 1834 bigrams and 10316 trigrams were trained for the phonotactic grammar using sub-phone units.

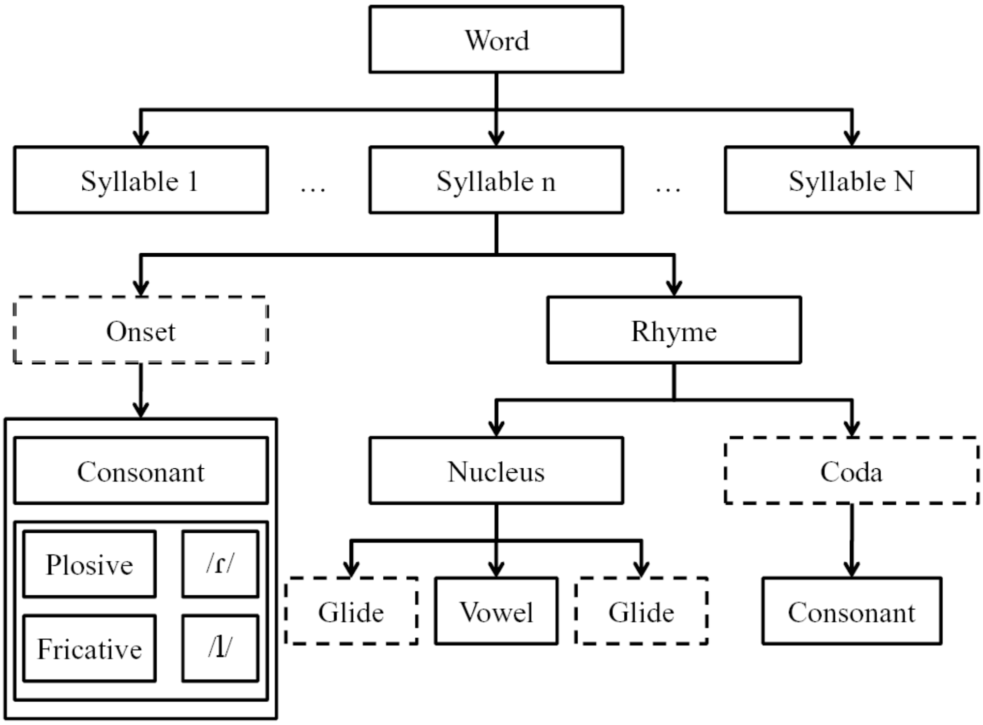


Figure 4.8: Rule-based phonotactic grammar

A rule-based grammar was created following the rules of creation of syllables in the Spanish language. This grammar followed the diagram in Figure 4.8, where a word was a sequence of syllables, whose structure was the traditional onset-rhyme structure, where the onset and coda were optional parts of the syllable. The onset of the syllable was every possible consonant or the cluster of a plosive or fricative with the vibrant /ɾ/ or the lateral /l/. The nucleus was a single vowel or a diphthong or triphtong created with the glides [j] or [w]. Finally, the coda was every possible consonant. This grammar fitted well to a big majority of the possible instances in Spanish, being more than enough for the words in the target task.

4.5.2 Results in the unimpaired speakers

The results with the reference speakers in Table 4.9 marked the good properties of the designed APD system. The results with the sub-phone TD models (8.48% PER) achieved a decent performance for this task, indicating the accuracy of the phonotactic recognizer used in this task. Another interesting conclusion was the similar performance achieved by the rule-based grammar in comparison with the trained data-driven grammar, marking the well-matching of the syllable creating rules to the language. Anyways, these results would require to be generalized to a more difficult task to evaluate this performance of the two models. Furthermore, the rule-based grammar was tested only with the the phone units due to the difficult inclusion of the context dependencies in the grammar that was the main feature of the sub-phone units.

Table 4.9: Baseline APD results for the unimpaired speakers (TI models)

PER results with TI models			PER results with TD models		
Phoneme models		Sub-phone models	Phoneme models		Sub-phone models
Rule-based	Stochastic	Stochastic	Rule-based	Stochastic	Stochastic
33.40%	32.55%	26.42%	16.24%	15.76%	8.48%

4.5.3 Results in the impaired speakers

Both TI and TD models were evaluated in these experiments with phoneme and sub-phone models, and the results are given in Tables 4.10 and 4.11. Results indicated a significant loss of performance for the impaired speakers compared to the reference ones, marking once again the strong influence of the speakers' disorders in the performance of the APD system as it did in the performance of the ASR system.

Table 4.10: Baseline APD results for the impaired speakers (TI models)

PER results with TI models							
	Phoneme		Sub-phone		Phoneme		Sub-phone
Speaker	Rules	Stochastic	Stochastic	Speaker	Rules	Stochastic	Stochastic
*Spk*01	52.31%	50.86%	53.68%	*Spk*02	60.70%	56.68%	61.99%
*Spk*03	58.90%	58.56%	62.93%	*Spk*04	47.69%	45.55%	53.34%
*Spk*05	75.00%	70.38%	67.12%	*Spk*06	51.37%	49.32%	49.14%
*Spk*07	58.99%	55.14%	57.79%	*Spk*08	59.85%	59.85%	63.44%
*Spk*09	63.27%	57.62%	58.30%	*Spk*10	61.30%	60.19%	65.50%
*Spk*11	45.55%	42.55%	39.73%	*Spk*12	79.54%	75.26%	75.86%
*Spk*13	72.09%	72.00%	72.52%	*Spk*14	58.56%	54.28%	58.90%
AVG	60.37%	57.73%	60.02%				

The gain for applying TD models was around 20% in all cases (sub-phone and phoneme models with rule-based and data-driven grammars); significantly smaller than the gain in ASR with TD models which was 30%. In this task, the mismatch between adult trained TI models and children trained TD models had less impact in the overall PER than it had in the WER, indicating that the speakers' disorders had a very stronger effect at the phoneme level than at the word level.

Table 4.11: Baseline APD results for the impaired speakers (TD models)

PER results with TD models

	Phoneme		Sub-phone		Phoneme		Sub-phone
Speaker	Rules	Stochastic	Stochastic	Speaker	Rules	Stochastic	Stochastic
*Spk*01	42.89%	41.18%	37.67%	*Spk*02	43.24%	42.04%	42.04%
*Spk*03	45.21%	45.29%	42.81%	*Spk*04	30.39%	29.62%	26.03%
*Spk*05	70.55%	68.16%	61.04%	*Spk*06	34.42%	33.82%	31.42%
*Spk*07	50.86%	49.74%	48.20%	*Spk*08	54.37%	54.02%	58.22%
*Spk*09	42.89%	41.27%	34.59%	*Spk*10	54.88%	53.51%	53.42%
*Spk*11	35.96%	34.67%	25.94%	*Spk*12	79.02%	76.71%	72.86%
*Spk*13	69.61%	67.38%	67.81%	*Spk*14	40.92%	38.44%	34.67%
AVG	49.66%	48.28%	45.48%				

4.6 Influence of the Acoustic and Lexical Disorders

The previous results all across this Chapter have shown the significant decrease in ASR and APD performance between unimpaired and impaired speakers. Considering that all the speakers in both groups were equally balanced in age and gender and that, the unimpaired and impaired groups were age-matched, it could be hypothesized that this difference in performance was due to the speech and language disorders that the impaired speakers were suffering. These impairments had acoustic and lexical effects on this speech, that, at the lexical level, were measured in terms of the phonetic mispronunciations made by these speakers and labeled by a set of human experts in Section 3.6. The presence of correlation between these mispronunciations and the ASR and APD results could assure the hypothesis of the influence of the disorders in the performance of both systems and somehow separate the acoustic and lexical effects of the disorders.

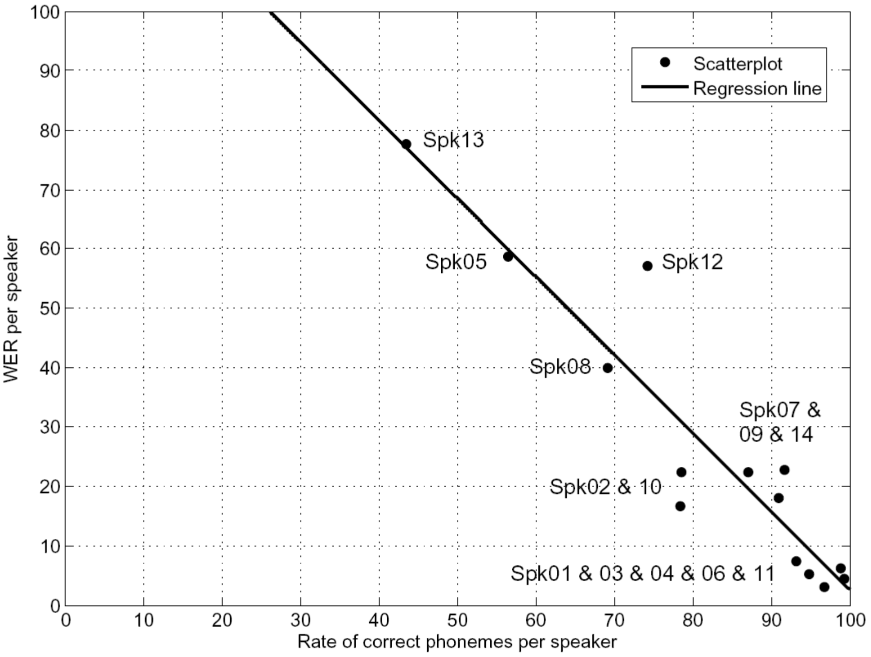


Figure 4.9: Correlation of the WER results with the percentage of correct phonemes

4.6.1 WER and mispronunciations

The possible relationship between the ASR results and the different impairments of the speakers arose when evaluating the separated results per speaker in terms of their rates of mispronunciations. In the best system described yet (word TD models), the scatterplot between the WER of each speaker against the rate of mispronunciations of that speaker is given in Figure 4.9. The same figure plots the regression linear function for this scatterplot ($y = 134.38 - 1.3173x$) whose r value was 0.938, indicating the existing correlation that could be detected at first sight from the scatterplot.

However, these results might have a dependency on the speakers' acoustical properties of their speech, so a further study was made to study in a speaker independent way how the number of mispronunciations per word affected the ASR results. In this case, a different WER was extracted for the words according to the number of phonemes within the word and the number of mispronounced phonemes in the word. These results can be seen in Table 4.12 for sub-phone TD models (where N/A indicates cases with no examples). As this Table indicated, words with no mispronounced phonemes (1,669, 52.29% of the total) achieved a WER (9.1%) closer to the results achieved in the same situation (word TD models) with the unimpaired speakers (2.04%), with the WER increasing as the number of mispronunciations increased for all word lengths.

Table 4.12: WER depending on the number of phonetic mispronunciations

Number of	Number of mispronounced phonemes in the word								
Phonemes	0	1	2	3	4	5	6	7	8
3	8.4%	30.5%	85.7%	N/A	N/A	N/A	N/A	N/A	N/A
4	10.9%	35.0%	66.7%	70.0%	N/A	N/A	N/A	N/A	N/A
5	11.1%	40.8%	57.6%	80.7%	100.0%	100.0%	N/A	N/A	N/A
6	5.4%	18.4%	50.0%	77.4%	85.7%	100.0%	N/A	N/A	N/A
7	1.7%	7.9%	34.6%	59.3%	91.7%	100.0%	100.0%	N/A	N/A
8	6.8%	14.6%	73.3%	88.2%	80.0%	96.2%	92.0%	100.0%	100.0%
9	0.0%	8.3%	33.3%	66.7%	66.7%	100.0%	100.0%	100.0%	100.0%
All	9.1%	30.7%	58.3%	75.6%	88.4%	97.6%	92.86%	100.0%	100.0%

With these results, it was seen how correctly pronounced words were consistently recognized with a high rate in all cases, while strongly mispronounced words were not recognized in nearly 100% of cases when half or more of the phonemes mispronounced. These results were considered for the ASR-based validation of word-level utterances in speech therapy tools like "Vocaliza", as it will be presented in Section 10.2.2. This tool based part of its evaluation in the performance of the ASR system, which has been shown to be highly dependent on the quality of the pronunciation in terms of the number of mispronunciations.

After this review, a more precise description of the different elements present in the disordered speech could be given. An approximation is shown in Figure 4.10 with the results obtained with sub-phone models. This Figure presents a line whose axis represents the values of WER: In the bottom right of the line, the 0.0% WER represents the ideal recognition system; the closest system is that in which the acoustic models are fully adapted to the task and the domain, which happened with the children speech TD models obtaining a 2.1% WER over children reference speech. In the top left of the line, the results with TI models over disordered children speech (36.7% from Table 4.2). From this point to the end of the line there are three effects that were separated with different experiments:

The training of TD models produced an improvement in the system, shifting the WER to a value of 28.2% (from Table 4.5). This 24.57% of difference was, hence, hypothesized to be caused by the *acoustic mismatch between adult and children speech.*

The evaluation of those words which were labeled with all their phonemes as correctly pronounced obtained a WER of 9.1% (from Table 4.12). This further 73.18% of difference was, hence, caused by the *lexical variants due to the speech and language disorders* in the speakers.

Finally, the remaining difference between this last value of 9.1% WER and the best possible system (2.1%) could be hypothesized to be caused by the *acoustic distortion due to the speech disorders* of the speakers.

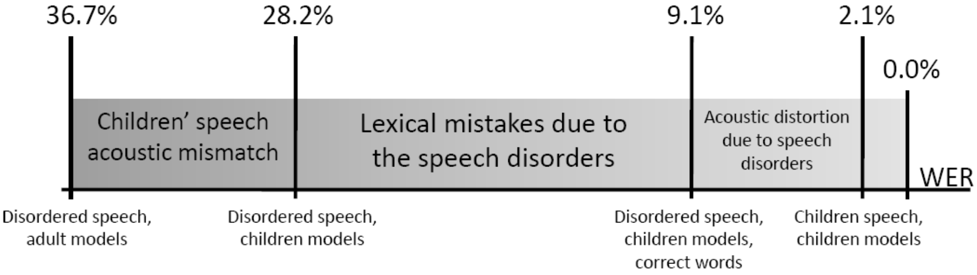

36.7%
28.2%
9.1%
2.1%
0.0%
Children' speech acoustic mismatch
Lexical mistakes due to the speech disorders
Acoustic distortion due to speech disorders
WER
Disordered speech, adult models
Disordered speech, children models
Disordered speech, children models, correct words
Children speech, children models

Figure 4.10: Influence of the different elements in speech over the WER

These hypotheses were just gross simplifications of the complicated processes of speech production in these impaired speakers, because the many interrelations between acoustic and lexical levels, as lexical mistakes could be produced by either top-level linguistic disorders or acoustic disorders and distortion, were not considered with the experiments that were done. Furthermore, there was always the underlying speaker variability in all the ASR experiments that could not be neglected, even when special care was taken to balance in age and gender the impaired speakers selection. Figure 4.10 just pointed all this possible effects in a simple and direct way.

4.6.2 PER and mispronunciations

A similar study could be made in terms of the PER of the proposed APD system seen in Section 4.5. While the PER achieved in the impaired speakers task clearly overpassed the PER of the unimpaired speakers, these results were obtained considering the canonical transcription as the ground truth of the speakers utterances. But the outcome of the human labeling showed than in 18% of the phonemes this was not accurate.

A different measure of the PER was considered: Instead of considering a recognition task, it was considered as a detection task. In this task, as it will be seen in the methods proposed in Chapter 8, the aim of the APD is to recognize correctly the actual pronounced phonemes and to reject the incorrectly pronounced phonemes (either deleting them from the output string or substituting them).

Table 4.13: FAR and FRR values for the outcome of the APD

	Phoneme models		Sub-phone models	
	FAR	FRR	FAR	FRR
TI	11.84%	43.33%	9.17%	51.17%
TD	12.26%	31.27%	10.52%	33.68%

The values in this task are shown on Table 4.13 for all the acoustic models evaluated in this Chapter in terms of False Acceptance Ratio (FAR) and FRR. These measures were defined as it has been traditionally used in detection problems as the rate of incorrect units falsely accepted and correct units falsely rejected in Equations 4.3 and 4.4.

$$FAR = \frac{\text{Incorrect units accepted as correct}}{\text{Incorrect units}} \tag{4.3}$$

$$FAR = \frac{\text{Correct units rejected as incorrect}}{\text{Correct units}} \tag{4.4}$$

The values of the False Rejection Ratio (FRR) are the values of PER for the correctly pronounced phonemes. These newly obtained values were closer to the results for the impaired speakers, but still significantly over [Saz et al., 2009d]. Only the data-driven grammars were evaluated, as the rule-based obtained a similar result for the evaluation done in phone models.

Once again, it was seen the different performance that ASR and APD systems could achieve when separating all types of existing data (lexically accurate or lexically incorrect). Furthermore, the proposed systems showed very high ability to discard incorrectly pronounced phonemes (FAR as low as 10%), although the rejection ratio of correct phonemes was too high to allow the use of the APD system as PV method without further help from other methods as it will be explain in further Chapters. It was also remarkable how TD models reduced the FRR (and PER) without increasing the FAR, indicating how these models were more accurate and fitted better to the children speech within the task.

Chapter 5

Acoustic Analysis of the Corpus

Come forth Lazarus!
And he came fifth and lost the job.

-James Joyce, *Ulysses*

In the two previous Chapters, a novel corpus with speech from children with disordered speech has been presented and the baseline results in recognition shown. A major degradation in the performance of ASR and APD systems has been shown, with a strong correlation with the grade of disorder of the speakers, at both the acoustic and lexical levels. The two following Chapters will try to further show the degradation of the acoustic and lexical properties within the impaired children' speech; with the current Chapter measuring the acoustic distortion in different acoustic features [Saz et al., 2006a, Saz et al., 2009g] of speech to understand the changes in the speech production due to the different speech impairments.

The following Chapter is organized as follows: Section 5.1 will review the most important acoustic and suprasegmental features that affect the production and perception of the Spanish vowels; with the method used for extracting these features also to be presented. In Sections 5.2 and 5.3, the values of these features for the set of speakers in the unimpaired and impaired corpus, respectively, will be presented. Finally, the comparative study of the values for both groups will be made in Section 5.3 and the possible degradations in impaired speech will be measured and studied.

5.1 Acoustic Features for Analysis

The selection of the acoustic features to study the quality of the speech from a given speaker is not an straightforward decision. Consonants are described by their articulatory features, and there is not a unique relation between acoustic and articulatory features that can help to define them acoustically [Neiberg et al., 2008]. Different works oriented to the detection of the articulatory features from purely the acoustic signal having shown these difficulties [Liu, 1996], as the only accurate way to detect articulatory features is by means of an electropalatograph. This way, for the study of the acoustic distortion introduced by the impaired speakers in their speech, it was decided to restrict the study to the 5 vowels of the Spanish language, which are described uniquely by their acoustic properties (mainly formants).

This imposed restriction limited the outcome of the acoustic analysis carried out in this Chapter, but it was assumable as the goal of the Chapter was just to obtain a brief outlook to the acoustic distortion hidden in the disordered speech from the impaired speakers, whose significance in the degradation of the ASR results was shown in Chapter 4. Hence, the selection of vowels to perform the acoustic analysis on the disordered speech in the corpus and measure the sources of acoustic degradation with relation to speech acquired from the age-matched individuals was reassured.

5.1.1 Acoustic features in the Spanish vowels

The Spanish language contains five vowels (/a/, /e/, /i/, /o/ and /u/) as seen in Appendix A. These vowels are defined by the position of the tongue in the mouth in a 2-dimensional way (up vs. down and front vs. back). There are two allophones of the /i/ and /u/ vowels acting like glides ([j] and [w], respectively) that, despite being close to the vowels, cannot be considered as vocalic sounds when they are unstressed and make the transition from or to a purely vocalic sound which is the nucleus in the syllable [Hualde, 2005]. Hence, these glides were never considered for analysis in this work. Apart from the information provided in the formants, other suprasegmental features affect the perception of vowels and were also studied to evaluate possible further difficulties within the production of speech by the impaired speakers.

Formant frequencies are the only acoustic feature needed to describe Spanish vowels, where these frequencies rely heavily on the articulatory properties of each vowel [Quilis, 1981]. As mentioned, the two main articulatory properties are the horizontal position of the tongue (defining palatal or front vowels vs. velar or back vowels) and the vertical position of the tongue (defining high vowels vs. low vowels). The vertical position of the tongue affects the first formant, while the horizontal position affects the second formant. Higher order formants like the third or fourth formants do not have a significant impact in Spanish vowels and were not considered in this work; moreover, tone does not have an impact either in the distinction of vowels.

According to this organization, Spanish has two high vowels (low first formant, 300-400 Hz): the velar /u/ (low second formant, 900 Hz) and the palatal /i/ (high second formant, 2300-2700 Hz), while only one low vowel (high first formant, 700-900 Hz) /a/ with a central position between palatal and velar (middle second formant, 1500-1700 Hz). Finally, two more vowels share a central-high position (high first formant, 500-600 Hz): the velar /o/ (low second formant, 1000-1200 Hz) and the palatal /e/ (high second formant, 2000-2400 Hz). These values of the formants are the approximate standard values for an adult healthy speaker [Martínez-Celdrán and Fernández-Planas, 2007], and some variations can be produced with age and gender, what made necessary the comparative study of the impaired speakers with age-matched unimpaired speakers.

There are three main acoustic features that affect the suprasegmental production in Spanish: Tone, intensity and duration. In isolated words like was the case in this analysis, these features mostly affect the distinct perception of stressed and unstressed vowels, although they do it in very different ways. Stress is considered in many phonetic theories as a binary feature that can be characterized as +stress or -stress, as perceived by the listener. In Spanish, each word presents an stressed syllable, marked over the vowel which is nucleus in that syllable. Several trends differ in which suprasegmental feature carries most of the stress information, although nowadays it is widely accepted that tone is the main carrier of stress [Fry, 1958], followed by intensity. Anyways, no categorical assertion can be made in this subject, as the main prosody of the sentence and other microprosodic features can affect this perception in different utterances, as well as in the different characterization of tone in each language.

Finally, duration also has an influence in the perception of stress, but it is very affected by the fact that every syllable has a canonic length, so the duration of a stressed vowel is only comparable to the duration of the same unstressed vowel when they are the nucleus of the same syllabic structure. Otherwise, no categorical conclusion can be made from the comparison of the duration of stressed and unstressed vowels. However, a correct control of speech production produces a correct control of phoneme duration and influences the perception of the speech by the listener.

5.1.2 Estimation of the acoustic features

The acoustic analysis carried out aimed to achieve a robust estimation of the four features concerned for study. State-of-the-art speech processing algorithms were implemented to estimate these values following the diagram on Figure 5.1 as also implemented in the speech therapy tool "PreLingua" for the improvement of phonatory controls in young children in Section 10.1.1. The speech processing was applied framewise (with a frame length of 25 ms. and a frame shift of 10 ms.) after obtaining the automated segmentation of the input speech via a Viterbi-based forced alignment. TD-HMMs used for ASR in Chapter 4 were used for the Viterbi alignment.

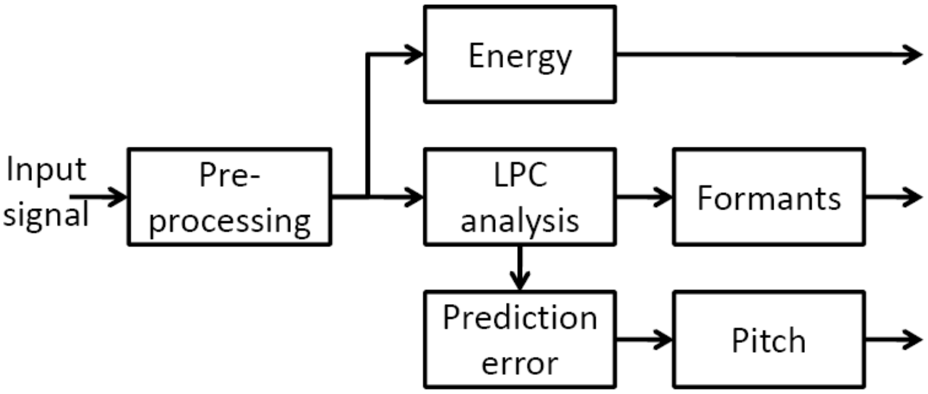


Figure 5.1: Estimation of pitch, energy and formants

An example of the outcome of the automated segmentation over one of the utterances of the reference speakers can be seen in Figure 5.2(a). The automated segmentation is initially based on the canonic transcription of every one of the utterances (isolated words) but, to avoid the pernicious effect of phoneme deletions in the pronunciations of the impaired speakers, the deleted phonemes (as perceived in the human labeling) were not fed as input into the automated segmentation, as shown in the example in Figure 5.2(b).

After segmentation, impaired speech was studied in two different groups: Correctly pronounced vowels and mispronounced vowels. This way, the intelligibility could be studied separately in the situations in which the human experts understood the vowel as correctly pronounced (lexical accuracy) and in the situation of perception of mispronunciations (lexical substitution).

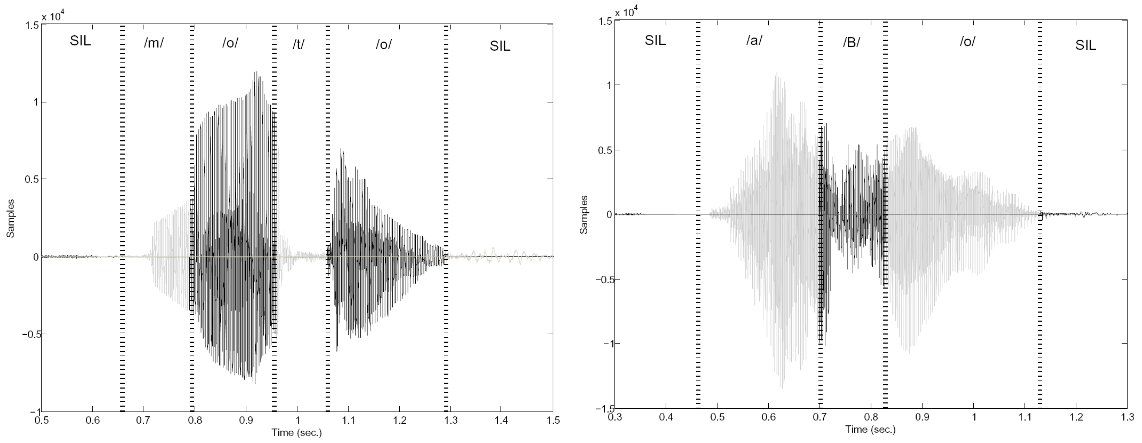
The feature estimation was carried out following the next steps: After signal pre-processing (DC offset, pre-emphasis and Hamming windowing), a LPC analysis was applied to every frame to extract the roots of the coefficients a_k in the 16-order speech prediction model in Equation 5.1.

$$H(z) = \frac{G}{1 - \sum_{k=1}^{16}(a_k z^{-k})} \tag{5.1}$$

Where the input signal $s(n)$ was estimated as $\hat{s}(n)$ using the time-domain impulsional response $h(n)$ associated to $H(z)$ in Equation 5.2, with $d(n)$ the glottal pulse signal.

$$\hat{s}(n) = h(n) * d(n) \tag{5.2}$$

The estimation of the formants used the 16 LPC coefficients a_k in the prediction model $H(z)$ and extracted the polynomial roots, each one of them associated to a formant frequency. The roots with the two higher absolute values corresponded to the first and second formants.



(a) Word "moto"

(b) Word "árbol" with deletions of /ɾ/ and /l/

Figure 5.2: Examples of segmentation via forced alignment

Tone estimation calculated the autocorrelation of the prediction error $e(n)$ given in Equation 5.3 and its autocorrelation $r(k)$ in Equation 5.4 with fr_l the value of frame length (25 ms. per frame).

$$e(n) = s(n) - \hat{s}(n) \tag{5.3}$$

$$r(k) = \sum_{n=0}^{fr_l} e(n)e(n-k) \tag{5.4}$$

The index k in which the autocorrelation had its maximum value outside from the area around the origin $r(0)$ was set as the pitch period (k_{pitch}) associated to the pitch frequency as in Equation 5.5 with an F_{sample} of 16 kHz.

$$F_{pitch} = \frac{F_{sample}}{k_{pitch}} \tag{5.5}$$

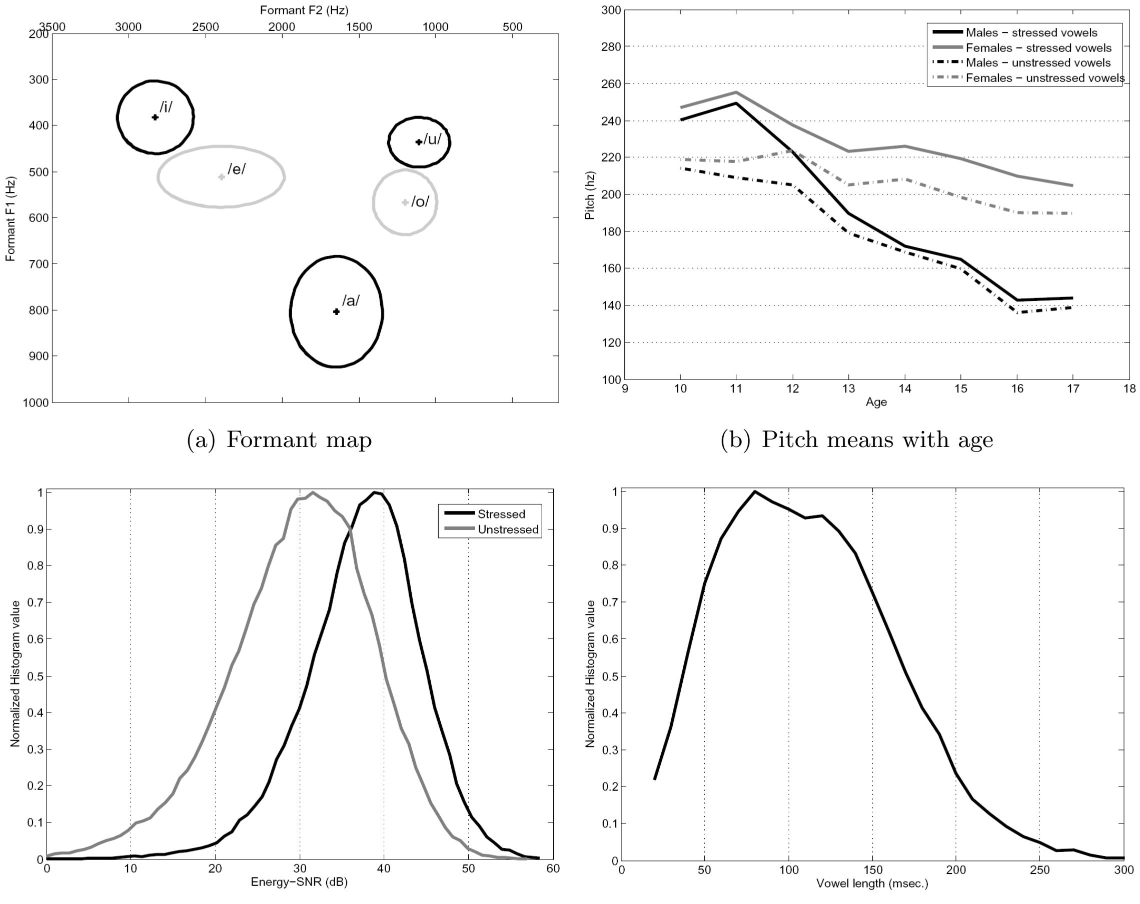
An estimation of the sonority value, as the ratio between the maximum value of autocorrelation and the autocorrelation in the origin (Equation 5.6), indicated if the frame was sonorant enough to be considered as a vowel and, hence, use the calculated pitch and formant values as correct. A high sonority ratio avoided the possibility of pitch and formant prediction mistakes (especially in vowel boundaries), although some correct frames might be rejected. Moreover, it intended to compute the formant values in the steady central part of the vowel to avoid possible coarticulation effects in them.

$$sonority = \frac{r(k_{pitch})}{r(0)} \tag{5.6}$$

For the intensity estimation, some arguments were considered prior to direct estimation. First, actual values of intensity (this is, sample values or directly computed frame energy) could not be considered into the study as it was not possible to reliably argue that input intensity during the recording process stayed steady through all different sessions, as the whole process of recordings from all the speakers took more than one year. However, it was reasonable to argue that SNR maintained steady values independently of the input volume since a close-talk microphone was used for the recordings.

This assumption was evaluated by the estimation of the background noise power level calculated for the corpus used in the work, whose mean value was 27.15 dB (7.22 dB of standard deviation) for the reference recordings and 27.07 dB (6.61 dB of standard deviation) for the recordings of the impaired speakers, which validated the hypothesis that noise level was directly related to intensity level and maintained similar and good properties through all the recordings. Hence, prior to energy estimation, average background noise power was calculated through all the frames considered as non-speech in the forced alignment. Afterwards, for each frame of the vowels, framewise energy was calculated and SNR obtained by subtracting the noise power in the utterance. The values of intensity provided through this Chapter will refer to these values of SNR as it was just described.

Duration calculation was done by estimating the length of the vowel in milliseconds, computing the number of frames assigned to each vowel in the forced alignment and then multiplying by the frame shift value of 10 ms. per frame. A threshold over the energy was applied to restrict the vowel boundaries and hence avoid the effect of coarticulation in the transitions to or from consonantal sounds. This threshold was pre-set to restrict boundary frames with low energy whose calculation of pitch and formants could be inaccurate.



(a) Formant map

(b) Pitch means with age

(c) Intensity histogram

(d) Duration histogram

Figure 5.3: Acoustic features in the reference speakers

5.2 Values of the Selected Features in Children' Speech

The reference 232 unimpaired young speakers were initially analyzed to determine the standard values of the formants and suprasegmental features under study in this work in a population in the same age than the impaired speakers. Some general assumptions were made concerning the statistical properties of the features studied in this work: First, the values of the formants were modeled as 2-dimension Gaussian distributions for each vowel. Values of pitch and energy were modeled like a Gaussian distribution separately for stressed and unstressed vowels; where values of pitch were also modeled independently for different speakers of age groups separately. Finally, the values of vowel duration were modeled as a Gaussian distribution.

Table 5.1: Formant statistics in the unimpaired speakers

	First formant				Second formant			
Vowel	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
/a/	803.8	120.2	0.33	-2.83	1645.2	299.3	0.23	-0.63
/e/	511.0	66.2	-0.45	-3.00	2396.6	410.7	0.49	0.13
/i/	382.0	78.8	1.10	-2.91	2825.3	247.3	-0.15	0.94
/o/	566.5	70.2	-0.40	-2.94	1198.6	206.3	1.49	3.16
/u/	436.3	53.9	-0.60	-2.98	1106.7	198.3	0.55	0.10

Hence, all these features were described in terms of their mean (μ), standard deviation (σ), skewness (γ_1) and excess Kurtosis (γ_2) values; where the values of γ_1 and γ_2 validated the Gaussian assumptions. Once assured the Gaussian properties in the reference speakers, μ and σ were the only statistics for the rest of the study. The graphical representation of these features for the unimpaired speakers is provided in Figure 5.3.

Table 5.2: Pitch statistics in the unimpaired speakers

	Stressed vowels				Unstressed vowels			
Group	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
Males 10yo	240.4	23.1	0.18	-0.14	214.2	25.8	1.34	5.38
Females 10yo	246.9	20.9	0.19	-0.23	218.8	23.6	0.99	3.74
Males 11yo	249.2	23.5	0.06	1.57	209.0	23.5	0.94	5.10
Females 11yo	255.2	28.3	0.20	-0.63	217.6	26.6	0.41	0.32
Males 12yo	222.9	31.7	-0.58	0.99	205.0	26.2	0.85	5.45
Females 12yo	237.4	22.8	-0.27	0.58	223.6	31.1	0.85	1.86
Males 13yo	189.9	31.4	0.10	0.86	179.1	30.1	1.39	7.62
Females 13yo	223.4	21.7	-0.21	0.28	205.0	23.5	0.83	3.07
Males 14yo	172.0	34.4	0.28	-1.22	169.0	27.3	0.18	-0.72
Females 14yo	226.0	17.6	0.12	1.82	208.3	22.0	1.51	7.29
Males 15yo	164.7	30.1	-0.07	-1.32	159.7	23.4	0.39	1.01
Females 15yo	219.2	19.1	0.16	1.25	198.3.2	24.4	1.59	7.35
Males 16yo	142.8	18.7	1.74	4.96	135.9	22.6	2.94	9.37
Females 16yo	210.0	20.9	0.08	-0.55	190.0	23.5	1.18	4.26
Males 17yo	143.8	25.0	1.24	1.75	138.9	22.9	2.37	6.65
Females 17yo	204.7	17.3	0.23	0.46	189.9	23.2	1.50	8.87

The results of the formant analysis in Table 5.1, plotted in Figure 5.3(a), obtained results in the range of the canonical values of the Spanish formants map, adapted to the known fact that those speakers with a higher pitch will also present higher formant values [Rodríguez and Lleida, 2009]. No normalization was applied to these formant values, as these values represented the standard formants for the population in the age of the impaired speakers. Dispersion of the values (measured as standard deviation) was in the reasonable margins for a set of different speakers like the one used in this work. The 5 vowels were correctly separable and distinguishable in the formant map, as could be expected from a group of healthy unimpaired speakers.

The pitch reference values were studied in terms of age and gender as in Table 5.2 and Figure 5.3(b). The difference in pitch between stressed and unstressed vowels was measured in around 20-30 Hz. for all cases, except for the older males where these differences got reduced in absolute value. As expected, values of pitch got reduced for the reference speakers as the age was increasing, with this reduction being especially noticeable for the male speakers. For the 17 years old speakers, the pitch values were close to the standard pitch values for reference healthy adults (140Hz for the male speakers and 190 Hz for the female speakers). The dispersion of the values (around 20-30 Hz. of standard deviation) was also between the expectable values.

Regarding the reference intensity values, a difference of 8 dBs in mean was measured between stressed and unstressed vowels, indicating the special emphasis that stressed vowels have on the production of speech. The histogram in Figure 5.3(c) and the values in Table 5.3 indicated the statistical properties of both values and the separability of them, as the dispersion of the values was kept in values around 8 dB of standard deviation.

Table 5.3: Energy statistics in the unimpaired speakers

Stressed vowels				Unstressed vowels			
μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
37.39	7.00	-0.88	4.25	29.90	8.90	-0.95	3.64

Finally, the duration of the vowels obtained from the reference speakers in Figure 5.3(d) and Table 5.4 gave a value of 111.8 msec. of duration per length with an standard deviation of 52.2 msec. These were normal values for the production of vowels in real speech, where every sound can take around 50-150 msec.

Table 5.4: Duration statistics in the unimpaired speakers

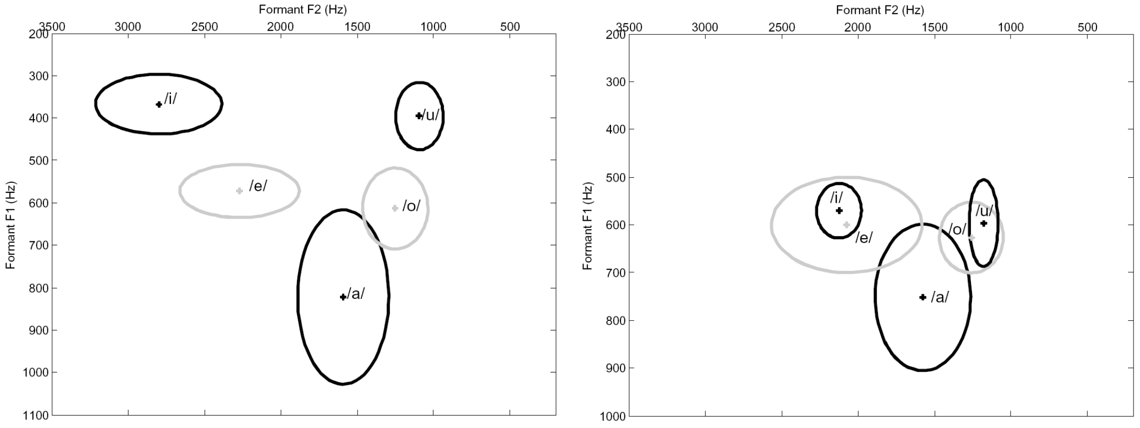
μ	σ	γ_1	γ_2
111.8	52.2	0.73	1.86

5.3 Results of the Analysis for the Impaired Speakers

After the study of the reference values for the four features subject of analysis in this Chapter, this Section brings the result of the analysis over the disordered speech with the same signal processing tools introduced previously and used over the reference speech. The analysis separated the correctly pronounced vowels from the mispronounced (substituted) vowels to understand how their acoustic properties differed when the speakers were correctly pronouncing the vowels or when producing a perceptible lexical mispronunication.

5.3.1 Formants in the disordered speech

The formant map for the 14 impaired speakers is shown on Figure 5.4. Figure 5.4(a) provides the formant map for the vowels perceived as correctly pronounced by the human labelers, with their statistics given in the first columns of Table 5.5. Two major effects were appreciated: First, the increase in the area of vowels /a/, /e/, /o/ and /u/ in the formant map in Figure 5.4(a), which was appreciated as an increase in the standard deviation of the formants in Table 5.5 when compared to the formants of the reference speakers in Table 5.1. And second, the approximation of vowels /a/, /e/ and /o/ towards the center of the formants map in Figure 5.4(a), also appreciated in the mean results in Table 5.5.



(a) Vowels labeled as correct

(b) Vowels labeled as mispronounced

Figure 5.4: Formant map in the impaired speakers

Concerning the results for the vowels perceived as mispronounced by the human labelers, given in Figure 5.4(b) and the second half of Table 5.5, it was appreciated a total confusion in the formants as expectable in this case where a mistake in the pronounced vowel was made by the speakers. In this case, all the formants were gathered around the middle of the formant map and the standard deviation increased in most of the cases, although the statistical values of standard deviation had to be taken cautiously because in some vowels there were not sufficient cases of mispronunciations for such a significant modeling.

Table 5.5: Formant statistics in the impaired speakers

	Correct vowels				Mispronounced vowels			
	First formant		Second formant		First formant		Second formant	
Vowel	μ	σ	μ	σ	μ	σ	μ	σ
/a/	822.1	205.6	1591.0	297.9	751.9	153.3	1575.8	313.0
/e/	571.9	62.6	2270.5	391.3	599.9	99.8	2074.0	492.8
/i/	366.9	70.5	2800.1	413.3	570.2	57.4	2124.5	146.6
/o/	613.2	95.5	1249.9	213.5	626.3	74.0	1258.1	208.8
/u/	395.0	79.7	1091.5	155.8	596.4	91.1	1178.8	94.8

In this case, what the speakers were really uttering was different from the canonical vowel to be expected and what was initially a given vowel had been uttered as a totally different sound. Another vowel, different from the canonical one, or a consonant, was uttered instead. As the labelers did not provide the alternative transcription to the uttered word, it was not possible to obtain the quality of the vowel production in those cases, but these results served secondarily as a validation of the reliable labels set by the human experts to correctly detect these mispronunciations.

5.3.2 Pitch in the disordered speech

The study of the pitch values for the impaired speakers was studied separately for each speaker; the high dependence of pitch in age and gender did not allow for the study of the full group of impaired speakers. These results for the impaired speakers are provided in Table 5.6 for correct and mispronounced vowels.

Table 5.6: Pitch statistics in the impaired speakers

	Correct vowels				Mispronounced vowels			
	Stressed		Unstressed		Stressed		Unstressed	
Speaker	μ	σ	μ	σ	μ	σ	μ	σ
Spk01	313.5	25.1	287.2	26.5	-	-	-	-
Spk02	299.9	18.9	258.8	26.2	288.9	28.3	274.5	28.8
Spk03	149.2	28.9	131.7	32.8	-	-	-	-
Spk04	264.6	21.1	226.8	20.4	248.5	16.2	224.3	10.9
Spk05	244.1	44.6	205.9	42.9	247.5	49.3	212.1	43.0
Spk06	140.2	31.9	130.1	33.3	-	-	-	-
Spk07	155.5	19.5	142.2	16.4	-	-	-	-
Spk08	207.9	19.9	178.4	28.5	211.0	23.5	185.8	21.2
Spk09	253.4	20.5	237.2	23.6	-	-	-	-
Spk10	247.4	30.1	217.5	30.4	259.9	39.4	218.2	16.9
Spk11	247.5	18.5	199.6	23.7	221.5	15.4	231.6	14.4
Spk12	156.7	12.3	145.6	36.0	154.1	8.3	145.1	16.5
Spk13	268.7	29.3	239.8	33.5	268.7	28.7	243.5	30.4
Spk14	265.3	21.9	223.4	27.1	233.1	12.2	204.7	7.8

It could be seen as impaired speakers kept a good control of these prosodic features: Speakers showed the ability to discriminate stressed vowels from unstressed vowels in similar ways to the reference speakers. However, some of the results of pitch values for mispronounced vowels could not be studied for speakers Spk01, Spk03, Spk06, Spk07 and Spk09, due to the lack of enough mispronounced vowels to obtain a significant statistical modeling. These speakers were the ones who produced a lower rate of mispronunciations and their influence in the study of how the acoustic features varied in the mispronounced phonemes was much smaller.

5.3.3 Energy in the disordered speech

Regarding the values of framewise energy (SNR as explained on Section 5.1), the average results for all the impaired speakers are given in Table 5.7. It was seen as the energy distinction between stressed and unstressed vowels was totally lost for the correctly pronounced vowels and was even reverted for mispronounced vowels, although this could be originated by some other causes not under study in this analysis. In general, impaired speakers tended to lower the intensity of their speech production, as unstressed vowels kept similar energy mean values than unstressed vowels by the unimpaired speakers.

Moreover, the dispersion of the energy framewise values increased for the impaired speakers, indicating that their speech production was more variable in terms of intensity than that of the impaired speakers.

Table 5.7: Energy statistics in the impaired speakers

Correct vowels				Mispronounced vowels			
Stressed		Unstressed		Stressed		Unstressed	
μ	σ	μ	σ	μ	σ	μ	σ
31.95	12.14	31.60	11.39	28.53	13.08	33.90	10.25

5.3.4 Duration in the disordered speech

The statistics for the results of the vowel length in the group of 14 impaired speakers are shown on Table 5.8. It could be seen that there was an increase in the average length of around 40 ms. for all vowels when compared to the reference speakers in Table 5.4, for both cases of correctly pronounced and mispronounced vowels. But what was more noticeable was the increase in standard deviation (double than the dispersion in the unimpaired speakers), which indicated the presence of vowels with an extremely variable length, meaning the existence of extremely long and extremely short vowels.

Table 5.8: Duration statistics in the impaired speakers

Correct vowels		Mispronounced vowels	
μ	σ	μ	σ
153.9	104.1	140.6	111.3

These results in the increasing of the dispersion of the duration, altogether with the increase in dispersion in intensity shown previously, indicated that the speech production was less steady for the disordered speech speakers.

5.4 Comparative Studies and Distortion Measurement

The results obtained in the previous Sections could give way to a discussion on several aspects of the vocalic production of impaired speakers. The computation of several measures of distance between statistical distributions like the Kullback-Leibler Divergence (KLD) and the Fisher's Ratio (FR) [Cover and Thomas, 1991] provided the required statistical analysis to understand the achieved results. These two measures are known to provide a good metric of the discriminative power of two different random variables. In this work, both helped to know the discriminative separation between vowels in the formant map and between stressed and unstressed vowels in terms of tone and intensity for both subsets of speakers.

For this work, it was considered the KLD definition for n-dimensional Gaussian distributions (2-dimensional in the case of formants and 1-dimensional in the other features). This definition, considered for two distributions $A = \aleph(\mu_A, \Sigma_A)$ and $B = \aleph(\mu_B, \Sigma_B)$ with μ_A and μ_B as mean vectors, Σ_A and Σ_B diagonal covariance matrices and n the dimension of the distributions, is given by Equation 5.7.

$$KL(A,B) = \sum_{i=0}^{n} (log(\frac{\Sigma_{A_i}}{\Sigma_{B_i}}) + \frac{(\mu_{A_i} - \mu_{B_i})^2}{\Sigma_{B_i}} + \frac{\Sigma_{A_i}}{\Sigma_{B_i}} \tag{5.7}$$

However, given this definition, the KLD had non-symmetric properties; this meant that $KL(A,B) \neq KL(B,A)$, so a symmetric Kullback-Leibler Divergence (sKLD) was defined in Equation 5.8.

$$sKLD(A,B) = \frac{KLD(A,B) + KLD(B,A)}{2} \tag{5.8}$$

Furthermore, the FR equation for the two n-dimensional Gaussian distributions $A = \aleph(\mu_A, \Sigma_A)$ and $B = \aleph(\mu_B, \Sigma_B)$ is given in Equation 5.9, and was symmetric by definition.

$$FR(A,B) = \sum_{i=0}^{n} \left(\frac{(\mu_{A_i} - \mu_{B_i})^2}{\Sigma_{A_i} + \Sigma_{B_i}} \right) \tag{5.9}$$

5.4.1 Formant degradation

Concerning the formants, there was an important decrease in sKLD and FR in the formant map between the vowels /a/, /e/ and /o/ in Table 5.9, while vowels /i/ and /u/ separate from the other 3 vowels, increasing their sKLD and FR in the formant map.

Table 5.9: Formant distances in unimpaired and impaired speakers

	Unimpaired		Impaired (correct)		Impaired (mispronounced)	
Vowels	sKLD	FR	sKLD	FR	sKLD	FR
/a/-/e/	18.59	6.74	17.43	3.26	4.26	1.42
/a/-/o/	12.04	4.42	6.49	1.71	5.06	1.26
/e/-/i/	5.91	2.37	11.36	5.59	5.61	0.08
/e/-/o/	22.89	7.13	16.34	5.37	11.17	2.37
/o/-/u/	4.99	2.27	7.42	3.43	2.18	0.18

To understand how this could affect the perception of vowels, it was studied the different rates of appearance of the vowels in the Spanish language, not only in this corpus, but in some other major text corpora in Spanish like the Europarl corpus [Koehn, 2005], which was seen in Section 4.5 as a good reference for phonotactic knowledge in Spanish, where the percentage of appearances of vowels was measured as 11.83% for /e/, 9.51% for /a/, 8.07% for /o/ and only 4.28% for /i/ and 1.74% for /u/ of all the Spanish phonemes. This way, the distance between /a/, /e/ and /o/ is the one that affects mostly in Spanish, and according to Table 5.9, the distances between these 3 vowels (/a/ vs. /e/, /a/ vs. /o/ and /e/ vs. /o/) are the ones that suffered a major reduction between unimpaired and impaired speakers.

This reduction in discriminative power became total in the situation in which the impaired speakers were making mispronunciations detected by the human experts. Once again, this result related to the fact that speakers were not uttering the vowel that they were expected to in the canonical transcription of the word.

5.4.2 Pitch degradation

Table 5.10 shows that there was no decrease in the weighted sKLD and FR in pitch between unimpaired speakers and impaired speakers. The values for the impaired speakers was obtained as the average of the sKLD and FR between stressed and unstressed vowels for each age group in Table 5.2; while the values of sKLD and FR for the impaired speakers was obtained as the average of the respective sKLD and FR values for each impaired speaker in Table 5.6.

These results were consistent with some previous works [Patel, 2002] where it was seen that heavily impaired speakers could control some prosodic features in their speech even when they lost their intelligibility in their vowel production. This control of the fundamental frequency was kept even in the case of mispronunciation of the vowel, which indicated that it was independent of the actual pronounced phoneme.

Table 5.10: Pitch distance of stressed-unstressed vowels in unimpaired and impaired speakers

Unimpaired		Impaired (correct)		Impaired (mispronounced)	
sKLD	FR	sKLD	FR	sKLD	FR
0.85	0.38	1.91	0.77	2.70	0.91

5.4.3 Energy degradation

It was in terms of energy (or intensity) where impaired speakers seemed to have bigger problems in the control of prosody and stress. There was a total reduction in the sKLD and the FR in the discriminative power between these two distributions, as seen in Table 5.11. For mispronounced vowels, even there was a contrary effect as unstressed vowels had more intensity that stressed vowels, although it was not significant.

Table 5.11: Energy distance of stressed-unstressed vowels in unimpaired and impaired speakers

Unimpaired		Impaired (correct)		Impaired (mispronounced)	
sKLD	FR	sKLD	FR	sKLD	FR
1.04	0.44	0.01	0.00	-	-

The origin of this total lack of stress discrimination by means of the energy, on the contrary to pitch, which kept properly the stress features, is due to the lower intensity in the stressed vowels and to the increase in the dispersion of the intensity values. This indicated, as mentioned before, that the speech production in the impaired was less steady than that of the unimpaired speakers.

5.4.4 Duration degradation

Finally, the study of the length of the production of vowels by the impaired speakers in Table 5.8 showed an effect of dispersion in the length of the vowels. This meant that vowels uttered by the impaired speakers were more often abnormally long or short. Actually, two separate effects could be appreciated; on one side, there was an effect of lengthening of the vowels (around 30% of increase in the mean values between Tables 5.4 and 5.8); and on the other side, the dramatic increase in standard deviation (double in the comparison between impaired and unimpaired speakers).

The increase in the duration might be due to hesitations or slowness in the impaired speakers to reassure their speech production in the cases in which they have difficulties for the correct pronunciation. The higher dispersion might indicate once again, as in the case of the intensity study, the inability of the impaired speakers to keep an steady oral production through different words and utterances.

Chapter 6

Lexical Analysis of the Corpus

> Logic, *n.* The art of thinking and reasoning in strict accordance with the limitations and incapacities of the human misunderstanding
>
> -Ambrose Bierce, *The Devil's Dictionary*

After evaluating and measuring the acoustic degradation on the disordered speech of the Alborada-I3A corpus, this Chapter digs into the evaluation of the lexical variants introduced in this speech by the speakers due to functional speech disorders. These lexical variants are analized as perceived by the human experts in Chapter 3 and are studied over the whole range of phonemes (consonants and vowels) to understand the possible origins and patterns of these lexical difficulties.

The Chapter is organized as follows: Section 6.1 will analyze the distribution of the impaired speakers' mispronunciations over the set of Spanish phonemes and the effect of some articulatory features like point and manner of articulation in these mispronunciations; finally it will measure the consistency of the speaker's pronunciation across different sessions. Posteriorly, in Section 6.2, the influence of the syllabic structure and context on these patterns of mispronunciations will be studied; and finally, in Section 6.3, the studies of the lexical degradation and variants introduced by these speakers will be compared to the processes and patterns in language acquisition of preliterate Spanish children.

6.1 Non-contextual Analysis of Phonetic Mispronunciations

The presentation of the words in the RFI in Section 3.4 showed how the 24 phonemes of Spanish language in Appendix A (23 with the reduction of /ɟ̞/ to /ʎ/ by yeísmo) appeared in different amounts in this set of words. The comparison of how the different phonemes were mispronounced with different rates by the impaired speakers could be seen as an initial step to understand the phonological and phonetic patterns of the mispronunciations.

Although it was seen that each speaker had a very different rate of mispronunciations; with some speakers producing up to 56% of mispronunciations, and others barely reaching the 1% of mispronunciations, all the analysis of the lexical variants in this Chapter was made totally speaker independent, gathering all the utterances from the 14 speakers. The possible loss of specificity due to the merge of speakers with such different characteristics was subdued to the fact that, by grouping as much data as possible the results were much more significant from a statistical point of view.

Tables 6.1 and 6.2 present the rate of correct, substituted and deleted phonemes for vowels and consonants separately. Vowels presented a very high rate of correctness (over 90% in average), except for the high vowels /i/ and /u/, in which higher rates of mispronunciations or deletions were observed. Consonants, on the contrary, suffered a more noticeable rate of mispronunciations for all cases, including some dramatic cases like /T/, /r/ and /rr/ with a correctness rate around 50% or lower. On the other side, the fricative consonant /x/ obtained a correctness rate over 90%. As it can be seen, there was no separation into sounds or allophones for the study in Tables 6.1 and 6.2, as this initial study was totally non-dependent on the actual context in which each studied phoneme appeared.

Table 6.1: Labeling rates for vowels
Percentage of correct, substituted and deleted phonemes per phoneme

Phoneme (SAMPA)	Number of Examples	Correct	Substituted	Deleted
/a/	3248	95.85%	2.37%	1.79%
/o/	2128	95.39%	3.53%	1.08%
/e/	1008	90.08%	5.75%	4.17%
/u/	504	84.33%	10.71%	4.96%
/i/	784	83.93%	7.91%	8.16%

Table 6.2: Labeling rates for consonants
Percentage of correct, substituted and deleted phonemes per phoneme

Phoneme (SAMPA)	Number of Examples	Correct	Substituted	Deleted
/x/	224	91.52%	6.70%	1.79%
/t/	784	87.76%	8.16%	4.08%
/p/	1064	87.41%	7.80%	4.79%
/J/	112	86.61%	7.14%	6.25%
/n/	672	84.08%	7.89%	8.04%
/L/	224	82.59%	15.18%	2.23%
/s/	560	79.46%	10.89%	9.64%
/b/	616	77.60%	17.21%	5.20%
/k/	504	76.79%	15.67%	7.54%
/f/	392	75.26%	19.64%	5.10%
/m/	448	69.42%	12.50%	18.08%
/d/	336	66.07%	22.62%	11.31%
/tS/	112	63.39%	35.71%	0.89%
/l/	952	61.45%	17.96%	20.59%
/g/	280	61.43%	26.07%	12.50%
/T/	224	58.48%	19.20%	22.32%
/r/	1008	50.79%	21.63%	27.58%
/rr/	168	35.12%	61.31%	3.57%

However, no conclusion could be extracted directly from these results. While some phonemes had a very high correctness rate (mainly vowels and some consonants like $/x/$), most of the phonemes were around a 60-80% correction rate, which did not indicate any special phonological difficulty. If a phonological difficulty was present for these speakers, it would have produced a total inability in the production of a certain phoneme or phonemes, as the speakers would have never acquired such sound and they would be totally unable of uttering it. Next step was, hence, to study the influence of possible difficulties due to the articulatory features of each consonantal sound, mainly the point and manner of articulation in consonants.

6.1.1 Influence of the point and manner of articulation

The point and manner of articulation are the two main features that describe the production of consonants, altogether with the feature of voiceness or unvoiceness, as it can be seen in Appendix A. These features describe the position of the tip of the tongue during the production of the sound and the mode in which the air flows through the vocal tract. Physiological or functional difficulties in the speech production may produce that the speaker presented a major difficulty in the production of those sounds whose articulation point and mode required the correct functioning of an element suffering an alteration (tongue, lips, etc).

The study of the rates of mispronunciation depending on the point of articulation was made and the results are presented in Table 6.3. The influence of the point of articulation did not seem relevant for the production of mispronunciations in these speakers. All points of articulation presented a rate of correctness around 75-80%, except for alveolar consonants, whose rate got decreased to 66.61%, accompanied by an increase to 18% of the deletion rate compared to the values achieved by the rest of the points of articulation.

Table 6.3: Labeling rate for different points of articulation
Percentage of correct, substituted and deleted phonemes

Point of articulation	Number of Examples	Correct	Substituted	Deleted
Lips	2128	80.78%	11.51%	7.71%
Teeth	1736	76.96%	14.98%	8.06%
Alveolar ridge	3360	66.61%	16.25%	18.81%
Palate	448	78.79%	18.30%	2.90%
Velum	1008	75.79%	16.57%	7.64%

On the contrary, the study on the manner of articulation produced some more conclusive results as seen in Table 6.4. Plosive, nasal and fricative consonants maintained a similar correctness rate around 75-80%, but laterals and vibrants got lowered down to 65% and 48% respectively, in which seemed like a major difficulty for these speakers.

Table 6.4: Labeling rates for different manners of articulation
Percentage of correct, substituted and deleted phonemes

Manner of articulation	Number of Examples	Correct	Substituted	Deleted
Plosives	3584	80.27%	13.42%	6.31%
Nasals	1232	78.98%	9.50%	11.53%
Fricatives	1512	75.86%	15.61%	8.53%
Laterals	1176	65.48%	17.43%	17.09%
Vibrants	1176	48.55%	27.30%	24.15%

From all the achieved results, very little conclusions could be extracted anyways. Neither the point or the manner of articulation seemed to have a major influence in how speakers produced their mispronunciations in their speech. The high mispronunciation rate of the phonemes /r/ and /rr/ indicated a possible difficulty in the production of the alveolar vibrant consonants, also shared by the lateral consonants /l/ and /ʎ/. These two features (vibrants and laterals) are known to be the most complicated for the acquisition of preliterate children or foreigners; but these problems were not explained by possible phonological problems in the speakers, as that would result in the total inability to utter them, which was not the case. Consequently, the influence of the phonetic context was seen as the main possibility that made the speakers commit mispronunciations in some situations of the phonemes, while uttering the same phoneme correctly in other situations.

6.1.2 Consistency in the speakers mispronunciations

A possible effect that might have led to the lack of conclusiveness results in the phonological study presented previously could be the possibility that the pronunciations given by the speakers of the same word in the different sessions were very different. This inconsistency in the speakers' pronunciation would make totally irrelevant any study of their production of mispronunciations as these might change easily from utterance to utterance of the same word, without relation to any stable pattern or origin.

The measure of this consistency or "rate of repeatability" was made directly at the phoneme level; for each phoneme of a given word in the corpus there were four utterances made by every speaker. This measure aimed to know how probable was to obtain a similar label in the phoneme of a word if the label of that phoneme in another utterance of the same word was known previously. The human labels given to every possible pair of phonemes were compared, and the number of times this comparison obtained a similar result were counted. The relationship between this value and the total number of comparisons (292 phonemes per speaker and 6 comparisons per phoneme: 1,752 total comparisons) provided finally the measurement of the "rate of repeatability" or consistency in the speakers' pronunciation.

Example 6.1.1 An example of the process for measuring the consistency in the speakers' pronunciation is shown on Figure 6.1. Considering the four utterances from a given speaker of the word 'árbol' in the four sessions, there are 6 possible comparisons of sessions: Session 1 vs. Session 2, Session 1 vs. Session 3, Session 1 vs. Session 4, Session 2 vs. Session 3, Session 2 vs. Session 4 and Session 3 vs. Session 4. In all of these cases, the labeling assigned to the 5 phonemes of the word are compared, giving a result for each case. The final value is obtained adding up the results from all cases. In the example, 24 out of the 30 possible phoneme comparisons obtain the same labeling, providing a final rate of repeatability or consistency of 80%.

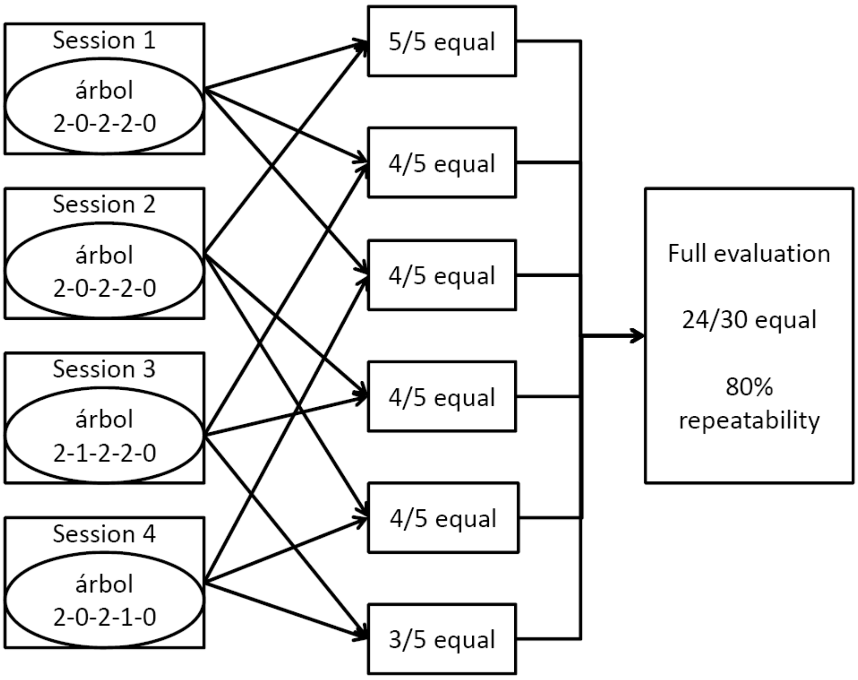


Figure 6.1: Measure of the rate of repeatability (consistency) in the impaired speakers

The complete results per speaker after this process are shown in Table 6.5. The mean result of 87.91% indicated that there was a very high consistency in the speakers' pronunciation and the pronunciation of a given word by a speaker could be predicted by listening to previous utterances of the same word. These results also validated the quality of the process of human labeling, as inconsistencies between the labels assigned to different utterances of the same word by a speaker might have been provoked by a very inconsistent labeling by the experts. It was to remember that each human expert was never handed more than 2 sessions of the same speaker (each speaker requested around 7 or 8 different experts), so this consistence in the labeling between sessions of the same speaker could not be due to the overexposure of these sessions to a same labeler.

Table 6.5: Consistency the impaired speakers' pronunciation

Speaker	Consistency	Speaker	Consistency
$Spk01$	98.00%	$Spk02$	87.10%
$Spk03$	91.04%	$Spk04$	95.77%
$Spk05$	72.26%	$Spk06$	98.74%
$Spk07$	88.47%	$Spk08$	81.62%
$Spk09$	94.29%	$Spk10$	83.85%
$Spk11$	94.29%	$Spk12$	77.96%
$Spk13$	73.06%	$Spk14$	94.24%
AVG	87.91%		

6.2 Context-based Analysis of Phonetic Mispronunciations

Once it was seen that no relevant conclusion could be obtained from the mispronunciation rates of the different phonemes in non-contextual positions and that the consistency with which the speakers tended to produce similar pronunciations of the same words in different days was high; it was considered to evaluate how the context influenced the production of mispronunciations by the impaired speakers.

However, the possible phonetic contexts of all the phonemes considering both the left and right neighboring phonemes were extremely elevated and that would lead to results with very low statistical significance due to the little amount of data for each context of study.

Hence, it was decided to study the context of the phoneme within the syllable that contained it. Syllables are full speech units with little influence from their neighbors (with this trend being more prominent in Spanish) and the study of the phoneme within the syllable can be considered independent from the rest of the word. Furthermore, the context within the syllable can be easily explained only in terms of the point and/or manner of articulation of the neighbor phonemes; as this will mark how complex is the coarticulation of the phonemes to be studied.

The complexity of the syllable was seen, hence, extremely relevant in the phonetic context within a syllable. Spanish presents a natural trend for CV syllabic structures (C: consonant, V: vowel), which leads to more complex structures with the presence of consonants in coda position CVC, diphthongs CVV, consonant clusters CCV or mixes of them in $CVVC$, $CCVV$ or $CCVVC$. Syllables with only one phoneme V can also appear in certain situations, as well as syllables with a triphtong VVV within their structure.

The rates of mispronunciation for all the speakers in terms of the length of the syllable in which the phoneme was included are presented in Table 6.6. There was a noticeable change in the rates in the pass from 2-phoneme syllables to 3-phoneme syllables, resulting in a decrease of around 7-10% of correct phonemes and an increase of 8-10% in deleted phonemes. This indicated that certain syllabic contexts were more difficult to pronounce, with disregard of the exact phonemes that appeared in them. The presence of codas, dyphthongs or consonant cluster appeared, hence, as a possible source of phonetic mispronunciations.

Table 6.6: Rate of mispronunciations per syllable length

Syllable length	Number of Examples	Correct	Mispronounced	Deleted
1-phoneme	112	88.39%	7.14%	4.46%
2-phoneme	5264	85.32%	10.47%	4.20%
3-phoneme	1680	77.50%	9.97%	12.54%
4-phoneme	168	75.30%	10.12%	14.58%

6.2.1 Influence of the syllable phonetic context

In the 57 words of the RFI which were the basis for the recording sessions with the impaired speakers, there was presence of vowels and consonants in different positions:

- Vowels as nucleus: /a/, /e/, /i/, /o/ and /u/.
- Vowels as glides in diphthongs: /i/ and /u/.
- Consonants as onset: /b/, /d/, /t/, /p/, /k/, /g/, /f/, /s/, /T/, /tS/, /m/, /n/, /J/, /l/, /L/, /r/ and /rr/.
- Consonants as codas: /m/, /n/, /T/, /s/, /r/ and /l/.
- Consonants as first parts of a consonant cluster: /b/, /g/, /p/, /k/, /f/ and /t/.
- Consonants as ending parts of a consonant cluster: /l/ and /r/.

Once seen this, phonemes that appeared in several syllabic positions were studied separately according to these different positions. The rates of correct, substituted and deleted phonemes were re-calculated for the phonemes in their different positions. As the standard syllable structure in Spanish is CV, the situation of onset and nucleus was considered as baseline for consonants and vowels, respectively, for the comparison of the other possibilities (consonant clusters, codas and diphthongs).

Table 6.7: Rate of mispronunciations for coda position

	Onset				Coda			
Phoneme	Number	Correct	Subst.	Deleted	Number	Correct	Subst.	Deleted
/m/	336	73.51%	16.07%	10.42%	112	57.14%	1.79%	41.07%
/n/	392	86.99%	10.20%	2.81%	280	80.00%	4.64%	15.36%
/T/	112	61.61%	27.68%	10.71%	112	55.36%	10.71%	33.93%
/s/	392	84.18%	11.48%	4.34%	168	68.45%	9.52%	22.02%
/r/	448	65.62%	24.33%	10.04%	224	35.27%	23.66%	41.07%
/l/	448	70.31%	23.88%	5.80%	168	54.17%	10.71%	35.12%

The first comparison was made among the 6 consonants who appeared on coda position, whose results are in Table 6.7. The number of appearances of the phonemes in each position is on the first column of the results. The number of these appearances was for all cases in the range of hundreds of examples for study (100-500). This was considered a very small number for what it could be seen as usual in language processing, so statistical significance of the given results could not be assured until a deeper statistical study was made. A certain trend could be seen, anyways, as the rate of correct phonemes decreased for all cases in coda position compared to onset position, with an increment of deletions and a slight decrease in substitutions.

For the 6 phonemes appearing in the initial position of a 2-phone consonant cluster, the labeling rates are shown on Table 6.8. Initially, there did not seem to be much change between the labeling rates in the two positions until a full statistical analysis was made, although the phoneme /g/ presented a noticeable increase in the rate of deletions when inserted at the beginning of this syllabic structure.

Table 6.8: Rate of mispronunciations for initial position of consonant cluster

	Onset				Consonant cluster			
Phoneme	Number	Correct	Subst.	Deleted	Number	Correct	Subst.	Deleted
/b/	448	79.91%	16.29%	3.80%	168	71.43%	19.64%	8.93%
/g/	168	69.64%	24.41%	5.95%	112	49.11%	28.57%	22.32%
/p/	896	87.72%	7.48%	4.80%	168	85.71%	9.52%	4.76%
/k/	448	77.90%	14.73%	7.37%	56	67.86%	23.21%	8.93%
/f/	280	75.36%	19.29%	5.36%	112	75.00%	20.54%	4.46%
/t/	728	87.64%	7.97%	4.40%	56	89.29%	10.71%	0.00%

For the 2 consonants appearing in the final position of a 2-phone consonant cluster, the re-estimated labeling rates are seen in Table 6.9. A noticeable increase in deleted phonemes in the consonant cluster position was seen, at the cost of a reduction of the correctly pronounced phonemes; even the rate of substituted phonemes was decreased.

Table 6.9: Rate of mispronunciations for ending position of consonant cluster

	Onset				Consonant cluster			
Phoneme	Number	Correct	Subst.	Deleted	Number	Correct	Subst.	Deleted
/r/	448	65.62%	24.33%	10.04%	336	41.37%	16.67%	41.96%
/l/	448	70.31%	23.88%	5.80%	336	53.27%	13.69%	33.04%

Finally, for the two vowels that acted like glides in diphthongs, the new separated labeling rates, shown in Table 6.10, marked an increase in substituted and deleted cases for the glides, decreasing the 90% correctness rate of the same vowels in nucleus position to a 70% correctness.

Table 6.10: Rate of mispronunciations for glide position

	Nucleus				Glide			
Phoneme	Number	Correct	Subst.	Deleted	Number	Correct	Subst.	Deleted
/i/	504	90.48%	6.35%	3.18%	280	72.14%	10.71%	17.14%
/u/	336	91.07%	7.44%	1.49%	168	70.83%	17.26%	11.90%

6.2.2 Statistical significance of the results

Once obtained the results of mispronunciations of the phonemes in different positions the questioning of the statistical significance of these results arose, as mentioned previously, due to the small number of cases for study. For this reason, z-test [Sprinthall, 2007] was used; this statistical method provides a measure, z, of how different are two groups for study according to their statistical properties in a full population. In this case, the full population were the realizations of a given phoneme, characterized by the rate of correct, substituted and deleted phonemes, and was divided in two groups: Phonemes in the baseline position and phonemes in the position of study; both groups with the labeling rates re-calculated previously. For these populations, the value z of the z-test was obtained as in Equation 6.1:

$$z = \frac{p_{position} - p_{baseline}}{\sqrt{p_{phoneme}(1 - p_{phoneme})(\frac{1}{n_{position}} + \frac{1}{n_{baseline}})}} \tag{6.1}$$

Where $p_{phoneme}$ is the rate with which the given phoneme was correctly pronounced, substituted or deleted, $p_{baseline}$ is that corresponding rate for the group of phonemes in the baseline position (onset for consonants, nucleus for vowels) and $p_{position}$ is the rate in the position of study (coda, consonantal cluster or glide). Finally, $n_{baseline}$ is the number of times the phoneme appeared in the baseline position and $n_{position}$ is the number of times the phoneme was seen in the position for evaluation.

The z-values for all cases studied in this Section are given in Table 6.11. Positive values indicated an increment in the rate (correct, substituted or deleted) from the baseline to the evaluation position, and negative a decrease. The decision of whether the two groups were different (baseline position vs. evaluation position) had to be made setting a confidence threshold that indicated statistical significance. Typically, confidence thresholds for the 95% or 99% of confidence interval are used in statistical studies; for these confidences, the z-value had to be over 1.96 or 2.575 in absolute value, respectively, to mark statistical significance

Table 6.11: Z-values for the analyzed phonemes

Phoneme	Position	Correct	Substituted	Deleted
/m/	Onset vs. Coda	-3.26 ↓	-3.96 ↓	7.30 ↑
/n/	Onset vs. Coda	-2.44 ↓	-2.64 ↓	5.90 ↑
/T/	Onset vs. Coda	-0.95 =	-3.22 ↓	4.17 ↑
/s/	Onset vs. Coda	-4.22 ↓	-0.68 =	6.50 ↑
/r/	Onset vs. Coda	-7.42 ↓	-0.20 =	8.48 ↑
/l/	Onset vs. Coda	-3.67 ↓	-3.79 ↓	8.01 ↑
/b/	Onset vs. Ini_Cluster	-2.25 ↓	0.98 =	2.56 ↑
/g/	Onset vs. Ini_Cluster	-3.46 ↓	0.78 =	4.06 ↑
/p/	Onset vs. Ini_Cluster	-0.72 =	0.91 =	-0.02 =
/k/	Onset vs. Ini_Cluster	-1.68 =	1.65 =	-0.41 =
/f/	Onset vs. Ini_Cluster	-0.07 =	0.28 =	-0.36 =
/t/	Onset vs. Ini_Cluster	0.36 =	0.72 =	-1.60 =
/r/	Onset vs. End_Cluster	-6.72 ↓	-2.58 ↓	9.90 ↑
/l/	Onset vs. End_Cluster	-4.86 ↓	-3.68 ↓	9.33 ↑
/i/	Nucleus vs. Glide	-6.70 ↓	2.17 ↑	6.84 ↑
/u/	Nucleus vs. Glide	-5.89 ↓	3.36 ↑	5.08 ↑

Table 6.11 also marks with ascendant or descendant arrows the situations in which statistical significance was seen with 95% confidence. A sign of equal close to the results in Table 6.11 marks that those results were not significant at that confidence interval.

The statistical significance study showed that most of the remarks seen during the calculation of the separated results were actually significant. All consonants suffered an statistical increase in their rate of deletions if inserted in coda position within a syllable when compared to the baseline onset position. This increase in deletions was accompanied by a decrease in correct phonemes (for /s/ and /r/), substituted phonemes (/T/) or both (/m/, /n/, /l/); indicating that the impaired speakers were reducing codas consistently independently of their ability to produce the same phoneme in the onset position or not.

Similar results occurred for consonants in the ending position of a consonant cluster. Phonemes /r/ and /l/ significantly increased their deletion rates when situated in this new syllabic context instead of in the baseline onset positions, at the expense of the rate of correct and substituted phonemes.

The results that did not show a sufficient statistical significance were the changes in phonemes acting as initial part of a consonant cluster. Only /b/ and /g/ showed a significant increase in deletions at the cost of a decrease in correct phonemes; but this trend was not even significant for phoneme /b/ with a 99% confidence interval. However, certain interrelations with the previous results could have been expected to happen, as if the second part of a consonant cluster is deleted (as it was shown to happen consistently) the structure *CCV* is reduced to *CV* and, hence, the consonant which started the consonant cluster is now acting as an onset, resulting in a better production for the speakers as it was shown for onsets.

The final study, for vowels in glide context, showed up the significant way in which /i/ and /u/ were mispronounced (substituted and deleted) when they were together with another vowel to create a diphthong. These results showed up to be significant too at the 99% confidence interval.

With all these results in hand, it was seen the dramatic influence of the syllabic context in the difficulty of the phoneme pronunciation for the impaired speakers. Complex syllabic structures were, this way, confirmed as a main source of mispronunciations in these speakers, due to the coarticulation effects and lexical difficulty of these syllables.

6.3 Relations with Language Acquisition Delays

The previous results showed a significant decrease in the ability of the impaired speakers to produce correctly consonants in coda position or consonant cluster and vowels in dypthongs. Several studies on the speech of preliterate children during the process of language acquisition showed some results that were seen as relevant to understand better the speech processes of the impaired speakers in the Alborada-I3A corpus.

The deep study in [Bosch-Galcerán, 2004] designed a vocabulary of 32 words for the assessment of the mispronunciations in a population of 293 children between 3 and 7 years, distributed in ages as in Table 6.12. The correctness in the pronunciation of these target speakers was evaluated by two different experts with with an agreement rate over 99%. All the children had Spanish as mother language and did not present any kind of voice or speech disorder or another impairment that could difficulty their speech production or acquisition.

Table 6.12: Distribution of ages in the study in [Bosch-Galcerán, 2004]

3 years old	4 years old	5 years old	6 years old	7 years old
50	70	64	54	55

The production of mistakes was studied to obtain the trends in similar situations to the ones proposed for the impaired speakers in this Chapter. Initially, [Bosch-Galcerán, 2004] studied the production of mistakes in the children according to the manner of articulation of the consonants (nasals, plosives, fricatives, laterals and vibrants). The results, summarized in Figure 6.2, showed up how the 3 year old children had problems with vibrants (more than 60% mispronunciations) with a 10-20% of mistakes in the rest of consonants. Older children produced more accurately all the phonemes; except for vibrants (/r/ and /rr/) and laterals (/l/ and /L/), that kept a 10-20% of mispronunciation rate in the 6-7 year old speakers.

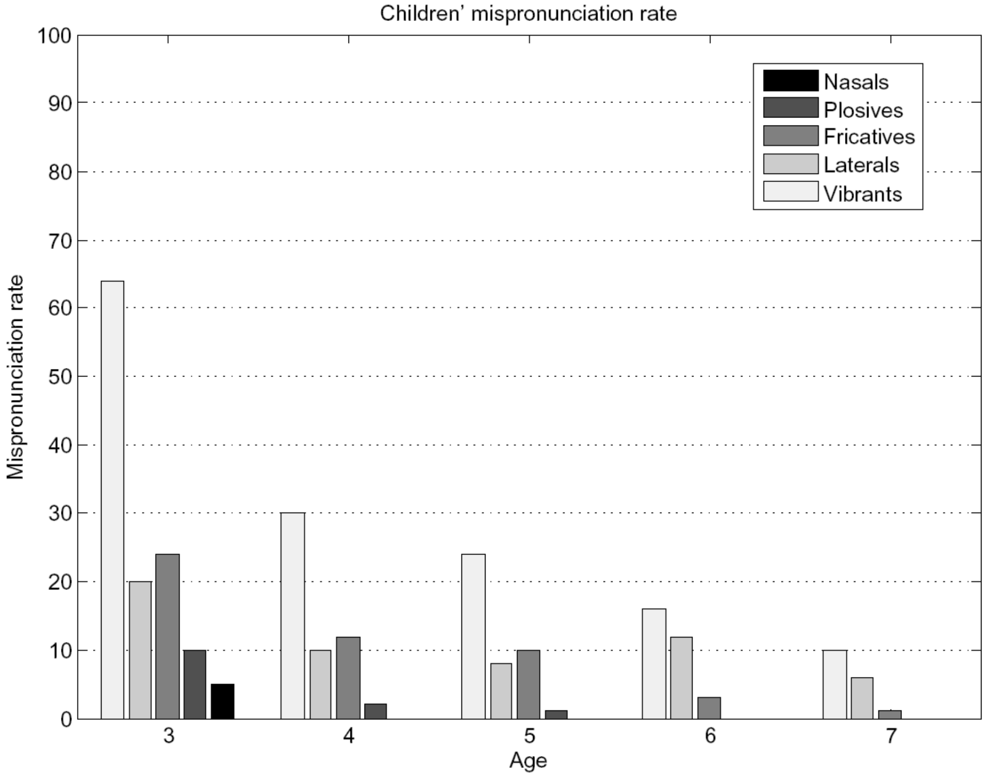


Figure 6.2: Mispronunciation rates in children' speech (results from [Bosch-Galcerán, 2004])

These results had many points of similarity to the non-context study of how impaired speakers produced mispronunciations (Table 6.4 in Section 6.1), in which vibrants had less than 50% of correctness and laterals stood by 65% of correctness, when the rest of the consonants reached 80% of correctly pronounced phonemes.

More similarities arose concerning three of the syllabic positions studied in this Chapter (coda, consonant cluster and dypthong), [Bosch-Galcerán, 2004] provided the set of results summarized in Table 6.13 of the mispronunciations of these structures for the speakers in the study. It showed how these syllabic structures were mispronounced in 20% of the cases by the 3-year old speakers; with these rates of mispronunciations being reduced in the older speakers to nearly disappear in the 7-year old children.

Table 6.13: Mispronunciation rates in children' speech (from [Bosch-Galcerán, 2004])

Position	3 years old	4 years old	5 years old	6 years old	7 years old
Coda	24.33%	11.83%	9.94%	4.54%	0.59%
Consonant cluster	21.9%	11.3%	9.1%	3.3%	1.3%
Diphthong	14%	4.33%	1%	0%	0%

The conclusion that arose from this comparison of results was that the lexical difficulties of the impaired speakers had relationship with problems of language acquisition that delayed the impaired children' speech to that of young preliterate children (3-5 years old). [Bosch-Galcerán, 2004] showed how vibrants and laterals and complex syllabic structures supposed a major difficulty for preliterate children and how these mistakes were corrected through the natural process of language acquisition.

All this discussion led this work to some areas far away from the initial objectives, as they merged with the knowledge in neurolinguistic [Caplan, 1987] and psycholinguistic sciences. This Chapter has shown how the speakers in the corpus, with different cognitive disabilities like Down's Syndrome, also have suffered severe delays in their speech and language acquisition. Different neurolinguistic trends aim to understand if language has some specific areas in the brain or not. In the case studied in this thesis, it was clear the relationship between cognitive impairment and speech impairment; but there have been many cases that have shown individuals with severe cognitive impairments who presented a perfect control over their speech and language.

Another element of study, in the humble approach of this Section to neurolinguistics, was how the speakers' mispronunciations resembled those of young preliterate children in an early age (3-4 years old). Although the speakers' mental age was lower than their physical age for many of the dairy life activities, as they had a high degree of dependence, in any case their overall mental age was as low as their linguistic age, an effect that proposed more discussion which, of course, could not be fulfill in this work.

However, we expect that this could be more deeply studied and analyzed in the future, and this knowledge can provide further elements for discussion to the work of all the speech technologies researchers dealing with this area.

Chapter 7

Acoustic-Lexical Adaptation for ASR of Disordered Speech

> He did not know that the new life would not be given to him for nothing, that he would have to pay dearly for it, that it would cost him great striving, great suffering.
>
> -Fyodor Dostoevsky, *Crime and Punishment*

Chapter 4 showed the dramatic reduction in ASR performance for young impaired speakers compared to their age-matched peers. Two causes for this loss have been shown in Chapters 5 and 6: On one hand, the acoustic distortion produced in the speakers' speech due to morphological and functional disorders affects the way their speech matches acoustic HMMs trained with unimpaired speech; while, on the other hand, the problems in speech and language acquisition in these speakers produced that their utterances did not match in many cases the canonical pronunciation of the words in the lexicon of the task.

This Chapter proposes speaker adaptation to avoid the loss of performance that these two factors produce in ASR. Speaker adaptation provides a framework to create an ASR system matched to the specific features of a target speaker. While, traditionally, speaker adaptation has referred to purely acoustic adaptation, the lexical variability in these speakers makes that two different approaches have to be taken in this Chapter: acoustic adaptation and lexical adaptation, studying their possible interrelations.

The Chapter is organized as follows: Sections 7.1 and 7.2 will provide the results of different strategies for acoustic and lexical speaker adaptation, respectively; while Section 7.3 will show how the ASR results improve with the joint use of both approaches (acoustic and lexical adaptation). Finally, Section 7.4 will discuss different issues in the interaction between both adaptation frameworks and the possible interrelations between them.

7.1 Acoustic Speaker Dependent Adaptation

Acoustic speaker adaptation aims to estimate the set of new HMM parameters that better fit to the speech characteristics of a given user from a set of utterances belonging to that speaker. The requirements for speaker adaptation are the set of input speech signals and their transcriptions. Different frameworks for speaker adaptation, MAP and MLLR, are presented and briefly discussed in Appendix B. When accurate transcriptions of the adaptation utterances are available and used for the creation of the SD models, it is called supervised adaptation; on the contrary, when it is required the use of a previous phase, like ASR or APD to obtain an estimate of the transcription, it is referred to as unsupervised adaptation.

The speaker adaptation carried out in the experimental part of this Chapter was based in a MAP adaptation [Gauvain and Lee, 1994] where the transition probabilities, the Gaussian weights and the Gaussian means in the HMMs were retrained and recalculated after 12 iterations of MAP. MAP adaptation was initially taken because it achieved fast and reliable performance in a situation like the one proposed in this Chapter, in which the task is controlled and all the words and units in the test data are seen on the train data. Further possibilities of MLLR adaptation [Legetter and Woodland, 1995] could be considered later if required.

A set of 4 leave-one-out experiments were designed as explained in Figure 7.1 with the 4 available sessions from each speaker: 4 models were trained with 3 different sessions from the speaker, and each model was used in the ASR of the utterances in the remaining sessions. The outputs of the four ASR systems were finally merged to calculate the final WER for the speaker. The initial model for all the acoustic speaker adaptation experiments was the SI-TD model trained in Section 4.3 with the 232 unimpaired children' speech.

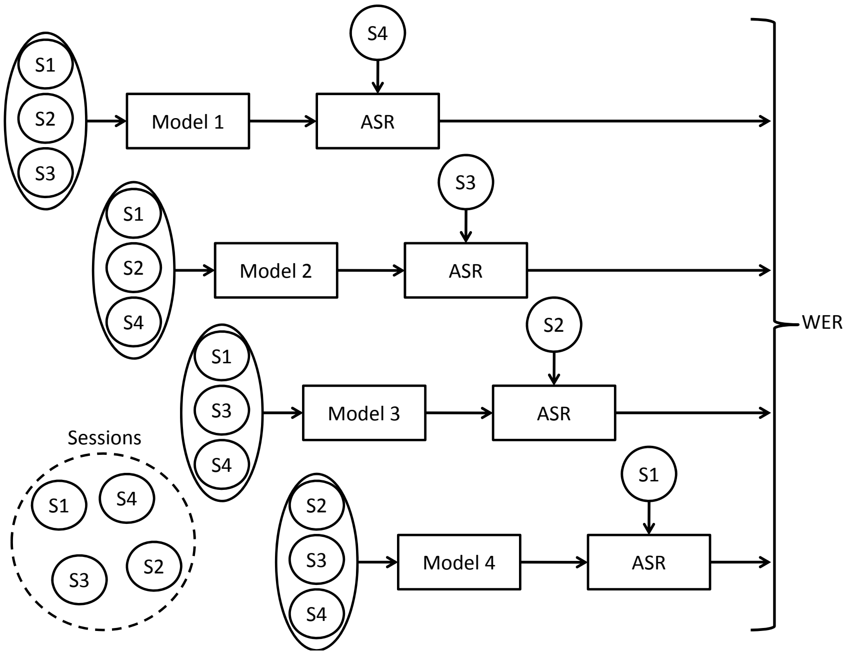


Figure 7.1: Design of the leave-one-out experiments

This process was carried out for all the adaptation strategies studied in this Chapter, including the estimation of different adapted models for the different HMM units used in Chapter 4 when possible (word, phone and sub-phone units). This way, all results were comparable as the process of adaptation was kept unchanged among experiments.

Three frameworks for acoustic speaker adaptation were evaluated at this point, based on Figure 7.2. The MAP adaptation system was fed with the train signals and with one out of the three possible set of transcriptions:

- The baseform transcription, corresponding to the prompts of the words used as adaptation data.
- The labels assigned by the human experts to the adaptation signals.
- The output of the SI-TD APD system seen in Section 4.5 over the adaptation data.

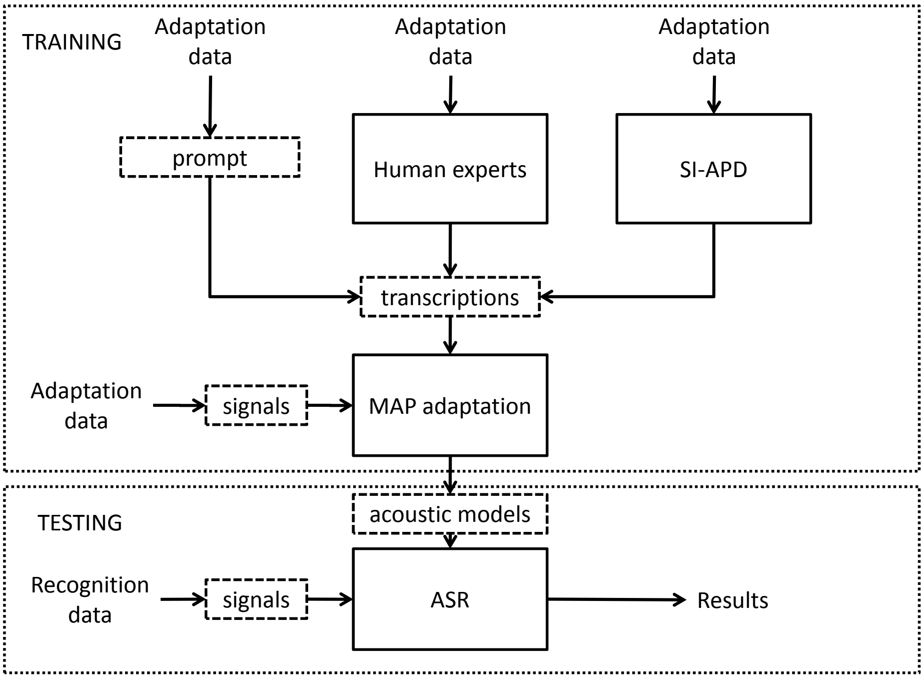


Figure 7.2: Acoustic adaptation frameworks

The initial strategy for speaker adaptation was to feed the canonical phonetic transcription of all the 57 words in the sessions to the adaptation system. The results with the SD models trained with this strategy are shown on Table 7.1. Results were available for word, phone and sub-phone models. The average WER for the three models (12.78%, 16.38% and 15.49%) supposed a significant improvement over the result with the TD models in Section 4.3. This gain achieved for the speaker adaptation was 50.56%, 48.75% and 45.07% for word, phone and sub-phone models respectively, with the improvement calculated over the TD results as in Equation 7.1. Once again, word models showed their outstanding performance over sub-phone and especially phone models.

$$Improvement(\%) = \frac{WER_{TD} - WER_{SD}}{WER_{TD}} * 100\% \tag{7.1}$$

Table 7.1: WER for ASR with acoustic models adapted to the canonical transcriptions

Speaker	word	phone	sub-phone	Speaker	word	phone	sub-phone
*Spk*01	2.19%	2.63%	2.63%	*Spk*02	7.02%	11.40%	10.96%
*Spk*03	2.19%	2.63%	2.19%	*Spk*04	2.19%	2.19%	2.19%
*Spk*05	40.35%	47.81%	44.74%	*Spk*06	1.75%	1.75%	1.75%
*Spk*07	6.14%	10.09%	7.02%	*Spk*08	19.74%	31.58%	28.51%
*Spk*09	10.96%	10.96%	10.96%	*Spk*10	8.77%	16.67%	12.72%
*Spk*11	2.19%	3.07%	2.19%	*Spk*12	11.84%	23.25%	17.54%
*Spk*13	58.77%	60.53%	66.67%	*Spk*14	4.82%	4.82%	6.58%
AVG	12.78%	16.38%	15.49%				

Evaluating every single speaker, 4 of them (*Spk*01, *Spk*03, *Spk*04 and *Spk*11) achieved WER results comparable to the results in unimpaired children with TD models. 5 more speakers obtained a WER around or lower than 10%, and the remaining 5 obtained higher values of WER, and as high as 60% in the case of *Spk*13. This showed up that for many of the speakers, simple acoustic adaptation was enough to provide a reliable performance in ASR and function similarly to their unimpaired peers.

However, there was a question that arose at the evaluation of these results, and it was how phonetic mispronunciations might alter the performance of the SI models trained with the baseform transcriptions. Although these transcriptions corresponded to the words that were prompted to the users; it was seen during the human labeling how the speakers produced a significant number of mispronunciations that altered these baseform transcriptions. As the real speaker utterances did not match the transcriptions fed to the adaptation system in the previous system, it could be expected that there was a certain effect associated to these mismatches between the utterances and the transcriptions.

The evaluation of this possible effect was made by retraining only with those phonemes within the utterances that were correctly pronounced according to the human experts. In this experiment, the human labels were also fed to the speaker adaptation system, that rejected the phonetic unit when it was labeled as substituted or deleted. Because of this phonetic approach, this experiment could not be run with word HMMs, because they retrained the word as a whole and did not allow for the phonetic separation.

Results with this strategy are presented in Table 7.2. The average WER for phone and sub-phone supposed a relative degradation of around 13% with respect to the models trained with the full canonical transcriptions.

Table 7.2: WER for ASR with acoustic models adapted to the labels of the human experts

Speaker	word	phone	sub-phone	Speaker	word	phone	sub-phone
*Spk*01	-	2.63%	2.63%	*Spk*02	-	14.91%	12.72%
*Spk*03	-	2.63%	1.75%	*Spk*04	-	2.63%	2.19%
*Spk*05	-	50.88%	46.05%	*Spk*06	-	1.75%	1.75%
*Spk*07	-	11.40%	9.21%	*Spk*08	-	34.21%	35.53%
*Spk*09	-	12.28%	11.40%	*Spk*10	-	19.74%	18.86%
*Spk*11	-	4.82%	3.07%	*Spk*12	-	27.19%	23.25%
*Spk*13	-	68.86%	70.61%	*Spk*14	-	7.02%	6.14%
AVG	-	18.64%	17.51%				

The loss of performance with these new models, which were supposedly more accurate that the previous models, could only be explained by the fact that the acoustic models might prefer inaccurate data which fitted better the inaccuracies that appeared during the recognition phase instead of more accurate data as decided by the labelers. Furthermore, the new strategy was using an 18% less of adaptation data than the previous one, because this was the rate of mispronounced phonemes that were not used for adaptation, and the adaptation system might have been very sensitive to this lower presence of data as it will be seen in Section 7.1.1.

The possibility of automating a way to adapt with only correct phonemes was then studied. In a realistic environment, it is not the case that it is possible to count with an expert, or a group or experts, who can decide whether some utterance is suitable or not for retraining of models, and automated decisions have to be made. At this moment, the APD system in Section 4.5 was used with this purpose. This system provided an estimation of the most likely sequence of phonemes, with a decent accuracy of prediction of mispronunciations, as seen in Section 4.6. The system was, hence, implemented and it used the phonetic decoded sequence as input in the MAP acoustic adaptation.

Results for this proposal are presented in Table 7.3. Although these results improved the baseline TD performance for phone units, there was a serious loss of gain compared to the previous proposals. It was seen, hence, that although the APD had some ability to detect mispronunciations with the same rate that the human labelers, as it was seen in Section 4.6, the recognition mistakes in APD were highly affecting the accuracy of the adapted models.

Table 7.3: WER for ASR with acoustic models adapted to the output of the APD

Speaker	word	phone	sub-phone	Speaker	word	phone	sub-phone
*Spk*01	-	12.28%	9.65%	*Spk*02	-	20.61%	22.37%
*Spk*03	-	10.09%	10.09%	*Spk*04	-	4.82%	3.07%
*Spk*05	-	55.26%	54.39%	*Spk*06	-	7.02%	4.83%
*Spk*07	-	28.07%	27.19%	*Spk*08	-	44.30%	45.61%
*Spk*09	-	20.61%	16.67%	*Spk*10	-	39.47%	35.09%
*Spk*11	-	15.35%	9.65%	*Spk*12	-	69.30%	67.54%
*Spk*13	-	73.68%	78.07%	*Spk*14	-	13.16%	14.04%
AVG	-	29.57%	28.45%				

7.1.1 Influence of the amount of adaptation data

All the results presented in this Section for the acoustic adaptation used 3 sessions for adaptation and 1 session for test. The amount of data available for adaptation is always a major issue in all adaptation frameworks [Saz et al., 2006b], as usually the performance of the systems grows up as the amount of data gets increased until a stable point is reached where more amount of the same data does not produce a significant change. The studies for increasing the performance of the speaker adapted models in cases of little adaptation data have been many during the years, as it is not always possible to count with large amounts of this data.

The framework adopted in this Chapter (171 utterances for adaptation, 57 for recognition) provided the biggest amount of data possible with only 4 sessions per speaker, but did not allowed for understanding how the system changed with different amounts of data. A small study on this subject was later possible with speakers *Spk*07 and *Spk*08, who counted with 8 sessions per speaker as seen in Chapter 3. A set of hold-out experiments were prepared with all the possible combinations of data for adaptation and recognition considering the available sessions. The organization of experiments is explained in Table 7.4, with all the number of experiments carried out for each case.

Table 7.4: Amount of Data for the Experiments

Number experiments	Adaptation Sessions (Utterances)	Recognition Sessions (Utterances)
8	1 (57)	7 (399)
28	2 (114)	6 (342)
56	3 (171)	5 (285)
70	4 (228)	4 (228)
56	5 (285)	3 (171)
28	6 (342)	2 (114)
8	7 (399)	1 (57)

Example 7.1.1 The hold-out procedure followed in these experiments can be seen with a simple example. For the case of re-training with 4 sessions and testing with 4 sessions, the possible

combinations of experiments are 70. The experiments include the case of re-training with sessions 1, 2, 3 and 4 and testing over sessions 5,6 , 7 and 8; as well as the contrary or other possible combinations like sessions 2, 4, 5 and 7 for training and 1, 3, 6 and 8 for testing. The order does not matter in these experiments, as it is the same to re-train with sessions 1, 2, 3 and 4 than with sessions 4, 3, 2 and 1. At the end, the possible experiments when re-training with a number n of sessions and testing with $8 - n$ sessions is obtained as the binomial coefficient that indicates the number of ways in which n elements can be chosen from a set of 8 elements: $\binom{8}{n}$.

The experiments were run over the purely acoustic adaptation with the baseform transcription fed to the adaptation MAP framework. The WER achieved over the 8 sessions from each speaker with the baseline TD sub-phone models was 23.46% for *Spk*07 and 47.81% for *Spk*08, similar to the results in Chapter 4 for the initial 4 sessions from each speaker (25.44% and 43.86% respectively), what indicated that this new task with the 4 extra sessions was well designed and the speech in the new sessions was similar to the speech initially recorded. The improvements achieved by the different adaptation frameworks are plotted separately for each speaker in Figure 7.3 in terms of the number of sessions used for adaptation.

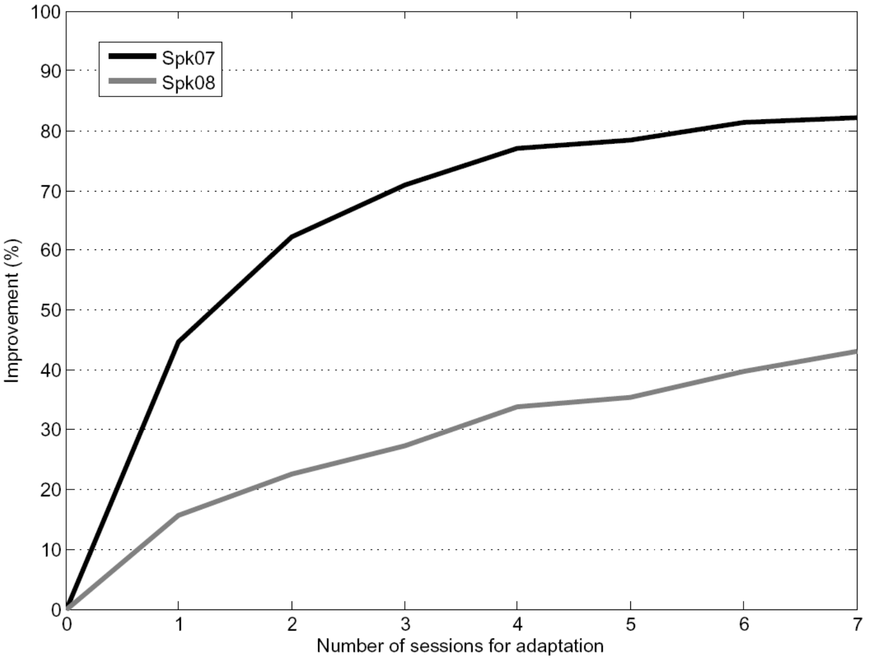


Figure 7.3: Influence of the amount of adaptation data in supervised acoustic adaptation

The relative marginal improvements achieved with every extra session added for adaptation are shown in Table 7.5. The marginal improvement for the adaptation with i sessions was the extra improvement achieved over the results with $i - 1$ sessions for adaptation as in Equation 7.2.

$$MarginalImp(\%) = \frac{Imp(i) - Imp(i-1)}{Imp(i-1)} * 100\% \tag{7.2}$$

The results in relative marginal improvement were slightly different for both studied speakers. *Spk*07 had a faster adaptation curve, and the fourth and following sessions could only provide less than 5% relative marginal improvements; for this speaker, hence, 3 sessions were enough for achieving a nearly optimal result. On the contrary, *Spk*08 had a slower adaptation curve, because the sixth and seventh sessions for adaptation provided a 12.32% and 8.76% of relative marginal improvement. The latent reason for these different reactions might be, once again, the different disorders and rates of mispronunciations for each speaker. This rate of mispronounced phonemes was significantly higher for *Spk*08 (30.82% vs. 12.93%), which meant that more inaccurate data was being fed to the adaptation and, hence, more amount of data was required for the accurate data to overwhelm the effect of the mispronounced phonemes.

Table 7.5: Relative marginal improvements for different amount of adaptation data

Speaker	1 Sess.	2 Sess.	3 Sess.	4 Sess.	5 Sess.	6 Sess.	7 Sess.
Spk07	-	39.28%	13.83%	8.54%	1.99%	3.59%	1.15%
Spk08	-	44.77%	20.83%	23.75%	4.52%	12.32%	8.76%

In all cases, the number of available sessions for adaptation with the rest of speakers (3 sessions) seemed sufficient to provide significant results close to the optimal results. Furthermore, the proposed adaptation framework was well-behaved as the improvement rates in Figure 7.3 raised steadily and no performance loss was found as the number of sessions grew up.

7.2 Lexical Speaker Dependent Adaptation

Works for lexical adaptation have been oriented traditionally to avoid the loss of performance in ASR on the use by dialectal speakers, non-native speakers and, like in this case, impaired speakers with speech disorders [Strik, 2001]. Rule-based strategies for lexical adaptation in Spanish have been used used for dealing with dialectal speech [Caballero et al., 2002], as this speech can be modeled according to the rules of pronunciations of the current dialect, although data-driven methods work well too [Caballero et al., 2004]. However, disordered speech presents very different lexical variants from one speaker to another and the use of pre-set rules could limit the performance of the system.

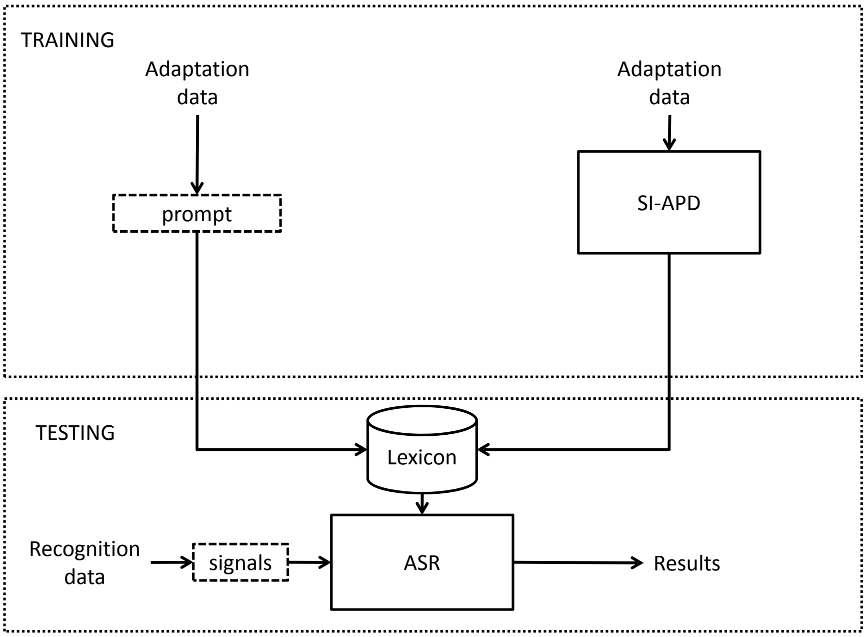


Figure 7.4: Framework for lexical adaptation

Data-driven methods, where new lexical variants of each word are estimated an introduced in the vocabulary will be, hence, mainly used for lexical adaptation. Although the initial works carried out towards lexical adaptation were focused on the possibilities that generic units based in the manner of articulation could offer for the substitution of mispronounced units [Saz et al., 2008c, Saz et al., 2008d], in a similar way to those works of [Fossler-Lussier et al., 2005]; they quickly turned to a dead line of work when it was seen that no improvement could be achieved with these techniques.

The proposed method for lexical adaptation in Figure 7.4 obtained the decoded phonetic sequence of the utterances included in the adaptation data via APD and added them to the lexicon with the canonical transcriptions of the words. This new lexicon was posteriorly used in the recognition phase. As 3 sessions were used in all cases for adaptation, the new lexicon included 3 new lexical variants for each word, which added to the baseform transcription supposed 4 possibilities per word. In the end, the 228 variants in the vocabulary competed freely against each other without any weighting on any of them.

The results with this approach for lexical adaptation are provided in Table 7.6. It was remarkable to see that, even when the acoustic models were kept unaltered, a significant improvement was achieved by directly using the proposed lexical adaptation framework. A possible drawback of the lexical adaptation was how it tended to achieve minor improvements (or even no improvement) in those speakers as *Spk*01 which do not produce so many lexical variants in their speech. In those cases, the novel expanded lexicons might be only adding confusability and lexical "noise", so a decision could be used to limit the new lexical variants that are introduced for each speaker in the dictionary.

Table 7.6: WER for ASR with lexical models expanded with the APD results

Speaker	word	phone	sub-phone	Speaker	word	phone	sub-phone
*Spk*01	-	10.53%	8.77%	*Spk*02	-	17.98%	16.23%
*Spk*03	-	18.86%	9.65%	*Spk*04	-	5.70%	3.07%
*Spk*05	-	55.70%	52.63%	*Spk*06	-	8.33%	5.70%
*Spk*07	-	24.56%	23.68%	*Spk*08	-	39.91%	33.33%
*Spk*09	-	17.98%	13.16%	*Spk*10	-	27.63%	32.89%
*Spk*11	-	9.21%	6.58%	*Spk*12	-	57.02%	50.44%
*Spk*13	-	64.04%	69.74%	*Spk*14	-	13.60%	14.91%
AVG	-	26.50%	24.34%				

7.2.1 Rule-based lexical adaptation

Data-driven lexical adaptation has already been argued to be the only assumable possibility for the lexical variants produced by the impaired speakers. However, Chapter 6 showed how the speech from these speakers presented a very similar trend with the speech of preliterate young children in the process of language acquisition. In [Bosch-Galcerán, 2004], a set of rules were proposed to characterize these speech processes at the lexical level for the preliterate children, which opened the gates for evaluating rule-based lexical adaptation in the impaired speakers. The rules were defined as follows:

- Rule 1: Velar consonants /k/ and /g/ are changed to labial consonants /t/ and /d/.
- Rule 2: Labial consonants /t/ and /d/ are changed to the velar consonants /k/ and /g/.
- Rule 3: Voiced plosives /b/, /d/ and /g/ are changed to the unvoiced plosives /p/, /t/ and /k/.

- Rule 4: Fricatives /f/, /s/ and /x/ are changed to the plosives with the same point of articulation /p/, /t/ and /k/.
- Rule 5: Plosives /p/, /t/ and /k/ are changed to the fricatives with the same point of articulation /f/, /s/ and /x/.
- Rule 6: Consonant /tS/ is changed to /s/ due to the loss of the plosive start in /tS/.
- Rule 7: Fricative dental /T/ is changed to the strident fricatives /f/ or /s/.
- Rule 8: Velar fricative /s/ is changed to the interdental fricative /T/.
- Rule 9: Vibrants /r/ and /rr/ are changed to the lateral /l/.
- Rule 10: Multiple vibrant /rr/ is changed to the simple vibrant /r/.

However, the direct creation of a new lexicon that included all these rules was not seen as useful for these speakers. The number of different variants in the lexicon grew up so much that the confusability in the lexicon was too high to produce good performance. Hence, it was decided to use a mixed data-driven/rule-based lexical adaptation. In this framework, the knowledge of the data came from the experts' labeling to detect the mispronunciations. The new lexicons for each speaker contained variants for each word according to the human labels: When a phoneme was marked as deleted, it was erased from the new variant; and when a phoneme was marked as substituted, the rules were used to determine the most plausible substituting phoneme in the variants. If no rule existed for a given phoneme, the canonical phoneme was kept.

Example 7.2.1 A simple example can be explained, with one of the utterances of the word "campana" (bell) by *Spk*05. Human labelers have marked the initial sound [k] as substituted and the inner [p] as deleted. According to the rules proposed in [Bosch-Galcerán, 2004], phoneme /k/ is usually substituted by [d], due to the rule 1 of the list, or by /x/, due to rule 5. Hence, for this speaker, the baseform transcription of "campana", [kampana] is accompanied by [damana] and [xamana] according to the detected substitution and deletion.

As this mixed technique was SD, all the leave-one-out experiments were replicated again, while keeping the SI acoustic models as in the previous experiments to provide only lexical adaptation. The results for this framework can be seen in Table 7.7.

Table 7.7: WER for ASR with lexical models based on rules and the labelers' data

Speaker	word	phone	sub-phone	Speaker	word	phone	sub-phone
*Spk*01	-	10.53%	10.09%	*Spk*02	-	18.42%	19.74%
*Spk*03	-	14.47%	7.89%	*Spk*04	-	4.82%	3.07%
*Spk*05	-	53.07%	45.61%	*Spk*06	-	7.46%	3.07%
*Spk*07	-	31.58%	23.68%	*Spk*08	-	44.74%	40.35%
*Spk*09	-	19.74%	18.42%	*Spk*10	-	33.77%	29.82%
*Spk*11	-	10.96%	8.33%	*Spk*12	-	73.25%	60.09%
*Spk*13	-	71.05%	74.56%	*Spk*14	-	19.74%	15.35%
AVG	-	29.54%	25.72%				

The results showed up some ability to improve performance over the SI lexicon results, but it showed worst performance than the fully data-driven proposal previously explained. Furthermore, it was never really proven that the speakers followed strictly those rules in [Bosch-Galcerán, 2004], as no real transcription from their utterances was available. Anyways, the improvement with the data-driven/rule-based SD lexicons indicated that in a certain way the rules could model the mispronunciations, which could be useful to works where no kind of previous data or a priori knowledge were available.

At this point, the use of rules for obtaining alternative lexical variants was abandoned to benefit the purely data-driven method. While the data-driven method was fully automatic, the rule-based proposal required experts' knowledge on the mispronunciations and a set of expert-defined rules; and, even with all of this, the later obtained better performance than the rule-based system, which made it not interesting at this point for further research in this thesis.

7.2.2 Influence of the amount of lexical adaptation data

In a similar way to how the influence of the amount of adaptation data was studied for acoustic adaptation, it was studied how having more lexical variants affected the performance of the ASR with a SD lexicon. The experiments prepared for this task were parallel to the ones carried out in Section 7.1.1 with the 8 sessions from Speakers *Spk*07 and *Spk*08, in terms of the number and organization of experiments.

In this case the number of lexical variants to be included in the SD dictionary ranged from 1, when only one session from the speaker went through the APD to obtain an alternative transcription, to 7, when 7 sessions followed this process. The number of total experiments for each possibility was the same as in Table 7.4. In the case with the smaller amount of data, there were 114 transcriptions competing in the ASR phase (2 for each word: one canonical and one learned from training data); and in the case with the larger amount of data, there were 456 transcriptions competing freely (8 for each word: one canonical and seven learned from training data). These values of competing transcriptions included those cases when a same transcription of a word was decoded twice from two different adaptation utterances, which reduced the number of effective competing transcriptions in the SD lexicon.

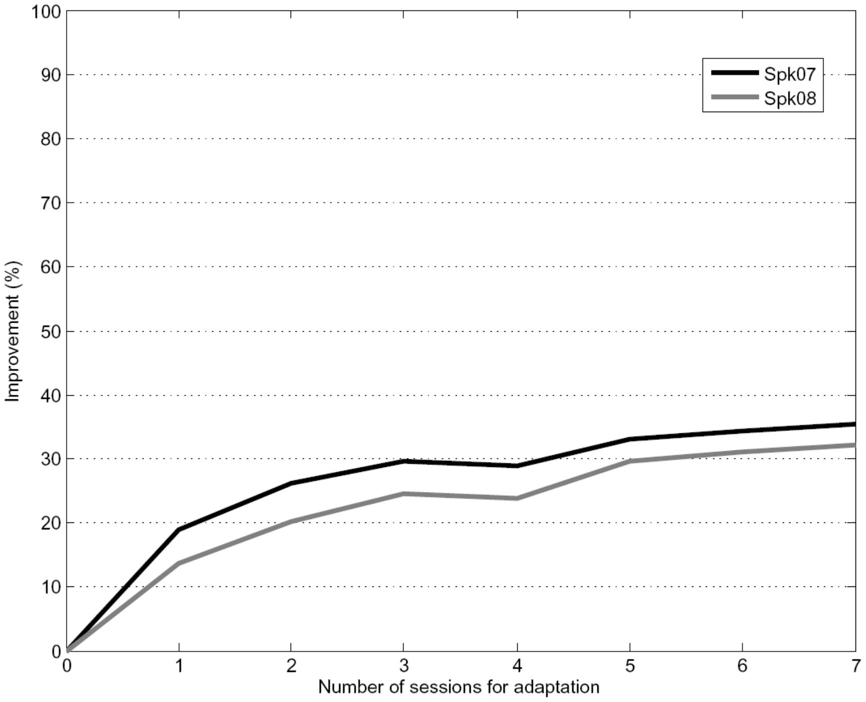


Figure 7.5: Influence of the amount of adaptation data in supervised lexical adaptation

The curve that shows the different improvements according to the amount of training data is in Figure 7.5, with the same process to extract the results than in Figure 7.3. The improvement was measured over the baseline TD WER of 23.46% for *Spk*07 and 47.81% for *Spk*08. Both speakers showed similar trends, achieving a 30-35% of improvement. This improvement was quite remarkable as it only made use of lexical adaptation. As it was seen in Table 7.6, lexical adaptation favored speakers with a larger number of mispronunciations, as they were those who requested more lexical variability in their SD dictionaries.

The results in terms of marginal improvement can be seen in Table 7.8. These values were not so steady in their growing, as it happened in the case in Section 7.1.1 with the acoustic adaptation. From this Table, also could be seen how the marginal improvement after the 5 session was minimal (5% or less), what indicated that the number of sessions for adaptation used for the rest of speakers was sufficient for this task to understand he performance of the proposed methods.

Table 7.8: Relative marginal improvements for different amount of lexical adaptation data

Speaker	1 Sess.	2 Sess.	3 Sess.	4 Sess.	5 Sess.	6 Sess.	7 Sess.
Spk07	-	37.84%	13.56%	-2.73%	14.79%	3.74%	3.48%
Spk08	-	46.81%	21.53%	-2.65%	23.88%	5.08%	3.16%

From these results, it could be seen how lexical adaptation could provide a major improvement in the long-term to these two speakers, who had an important prevalence of pronunciation mistakes.

7.3 Mixed Acoustic-Lexical Adaptation

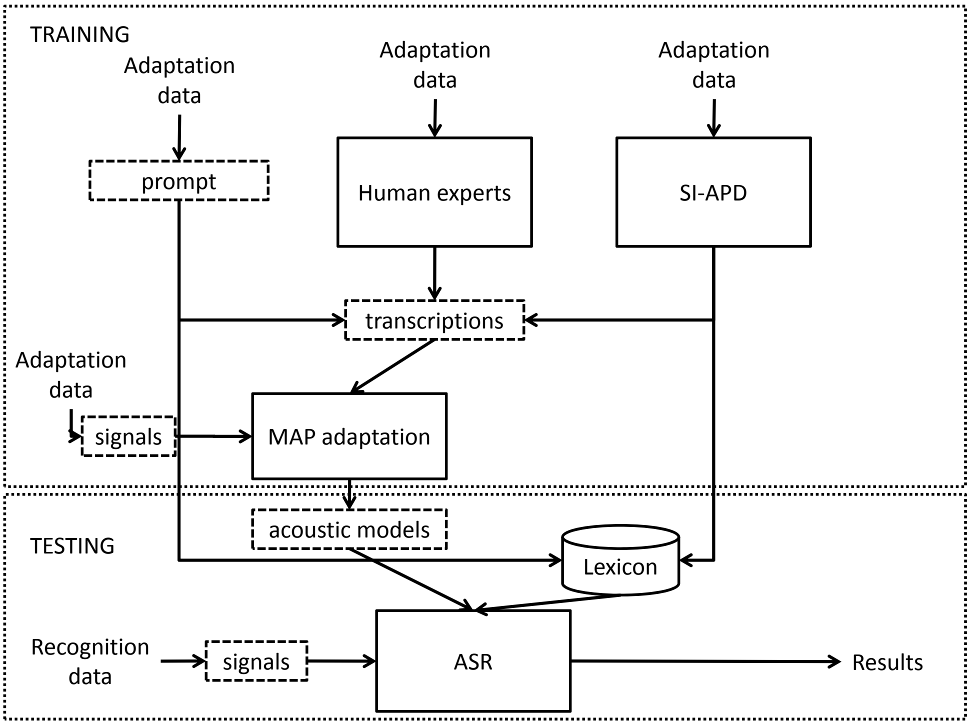


Figure 7.6: Acoustic-lexical adaptation frameworks

Finally, the combination of acoustic and lexical adaptation was evaluated [Saz et al., 2009c] to understand the possibilities that both approaches could provide when working jointly. This supposed the use of the three types of acoustic SD models trained in Section 7.1 with the data-driven SD lexicons obtained by APD in Section 7.2, as seen in Figure 7.6.

In the first case of acoustic models adapted to the baseform transcription and lexicon expanded to the APD outcome, the results in Table 7.9 for phone and sub-phone models (as word models do not accept lexical adaptation) showed a similar performance to the case of only acoustic adaptation in Table 7.1, although a certain loss in gain was observed for the average results obtained with the combined approach.

Table 7.9: WER in ASR with acoustic adaptation to baseforms and lexical adaptation

Speaker	word	phone	sub-phone	Speaker	word	phone	sub-phone
*Spk*01	-	2.63%	2.63%	*Spk*02	-	12.72%	12.72%
*Spk*03	-	3.95%	3.07%	*Spk*04	-	4.39%	2.63%
*Spk*05	-	46.05%	43.42%	*Spk*06	-	2.63%	2.19%
*Spk*07	-	11.40%	10.09%	*Spk*08	-	27.63%	29.39%
*Spk*09	-	9.65%	7.02%	*Spk*10	-	17.11%	16.23%
*Spk*11	-	3.51%	1.75%	*Spk*12	-	30.26%	28.51%
*Spk*13	-	62.28%	63.16%	*Spk*14	-	6.58%	6.14%
AVG	-	17.20%	16.35%				

Posteriorly, the results with acoustic models only trained with those units labeled as correct by the human experts and lexicons expanded with the help of the APD are shown in Table 7.10, improving the results in Table 7.2. The results with this approach did not have a significant difference with those recently shown in which the acoustic models were trained to the baseform transcription and lexical adaptation was further included. This showed up, as the initial results with only acoustic adaptation, that there was not much difference in limiting the amount of adaptation data as the units were correctly selected.

Table 7.10: WER in ASR with acoustic adaptation to human experts and lexical adaptation

Speaker	word	phone	sub-phone	Speaker	word	phone	sub-phone
*Spk*01	-	2.63%	2.63%	*Spk*02	-	13.60%	13.60%
*Spk*03	-	3.95%	2.63%	*Spk*04	-	4.39%	2.63%
*Spk*05	-	46.05%	46.05%	*Spk*06	-	2.63%	2.19%
*Spk*07	-	12.72%	10.96%	*Spk*08	-	31.58%	30.26%
*Spk*09	-	12.72%	9.21%	*Spk*10	-	15.79%	18.86%
*Spk*11	-	3.51%	2.63%	*Spk*12	-	33.77%	30.70%
*Spk*13	-	59.65%	63.60%	*Spk*14	-	5.70%	6.14%
AVG	-	17.76%	17.29%				

Finally, the results with acoustic models trained with the transcriptions as obtained by the APD system and the lexicons expanded with the same APD transcriptions are shown in Table 7.11, improving those of Table 7.3. These results were the ones that achieved a bigger improvement from including the lexical adaptation compared to the two previous ones.

As it will furtherer explained in the next Section, the lexical adaptation proposed was matching totally in this case the transcriptions that the acoustic models learned in adaptation. In this case, hence, lexical adaptation was providing the acoustic models a way to match the utterances from the speaker that the baseform transcriptions could not provide initially.

Table 7.11: WER for ASR with acoustic adaptation to APD and lexical adaptation

Speaker	word	phone	sub-phone	Speaker	word	phone	sub-phone
*Spk*01	-	7.46%	8.33%	*Spk*02	-	15.79%	17.74%
*Spk*03	-	9.21%	18.86%	*Spk*04	-	4.82%	2.19%
*Spk*05	-	51.75%	52.63%	*Spk*06	-	6.14%	5.70%
*Spk*07	-	17.54%	19.74%	*Spk*08	-	36.40%	34.65%
*Spk*09	-	13.16%	10.53%	*Spk*10	-	24.56%	27.63%
*Spk*11	-	8.33%	3.95%	*Spk*12	-	56.14%	42.54%
*Spk*13	-	61.84%	63.60%	*Spk*14	-	13.45%	10.09%
AVG	-	23.33%	22.71%				

7.4 Discussion on the Joint Use of Acoustic and Lexical Adaptation

Several frameworks for acoustic and/or lexical speaker adaptation have been presented. This Section aims to bring together all the results achieved to provide in a quick glimpse a comparison among the performances of all of them. With this comparison, the interrelations between acoustic and lexical adaptation were considered and studied.

The relative improvements achieved by all the proposed methods for phone and sub-phone models with respect to the TD baseline results in Section 4.3 (31.96% for phone models and 28.20% for sub-phone models) are presented in Table 7.12. In horizontal, the different possibilities in acoustic adaptation are presented (from the SI models to the three frameworks evaluated); and, in vertical, the two possibilities for lexical adaptation (no adaptation and data-driven) are reviewed.

Table 7.12: Relative improvements achieved by the different proposed methods

	Acoustic adaptation							
	No adaptation		Baseform		Human		APD	
Lexical adapt.	Phone	Sub-ph.	Phone	Sub-ph.	Phone	Sub-ph.	Phone	Sub-ph.
No adapt.	-	-	48.75%	45.07%	41.68%	37.91%	7.48%	-0.89%
APD	17.08%	13.69%	46.18%	42.02%	44.43%	38.69%	27.23%	19.47%

First point of study was the improvement that lexical adaptation alone provided (17% and 14% for phone and sub-phone units respectively). This improvement could be considered quite remarkable as the lexical variants were obtained in a fully automated way, and the acoustic modeling (which was kept unaltered in this case) has a major impact in the performance of all ASR systems.

Another element of interest was to see how the use of lexical adaptation improved the results achieved by the acoustic models trained with the human labels and the APD transcription, but it resulted in a loss of performance when acoustic models were adapted to the baseform transcriptions. The cause for this could be explained in terms of the match/mismatch between acoustic and lexical adaptation, and it pointed out the need of matching carefully both approaches to obtain the best results out of them.

The bigger impact of the lexical adaptation, anyways, occurred when the acoustic models were retrained according to the same APD transcriptions that were included as new lexical variants in the vocabulary. In this case, the match between acoustic and lexical adaptation was total and, although the absolute performance compared to other methods was lower, it pointed out that a full matching between both elements (acoustic and lexical) could boost the ASR performance if a better detection of mispronunciations would have been used.

Chapter 8

Confidence Measuring for Detection of Mispronunciations

> But that was not all. Travelers who ventured into that wild country found themselves assailed, it was said, by countless terrors which would make even the stoutest hearts to tremble.
>
> -Jan Potocki, *The Manuscript Found in Saragossa*

At this point, it has been seen how the acoustic and lexical distortion in the disordered speech has produced a dramatic effect in the loss of performance of the baseline ASR system. Adaptation strategies at both acoustic and lexical levels have been tested to avoid these effects, with the interrelations between both of them studied. From these experiments, it was seen how the correct knowledge of the phonetic mispronunciations made by the speakers was necessary for a correct acoustic-lexical modeling. The automatic detection of these mispronunciations is, hence, necessary for this task and, furthermore, for the development of automated CALL tools, which was set as another of the goals of this thesis.

This Chapter focuses on scoring methods for assessing the speech quality from the impaired speakers and detecting their phonetic mispronunciations. The ground-truth reference for the detection of these mispronunciations is the human labeling seen in Chapter 3. The different techniques will be oriented to avoid the effects of speaker variability in the PV system via score normalization and speaker adaptation.

The Chapter is organized as follows: Section 8.1 will introduce the problem of phoneme verification and how it resembles the traditional speaker verification problem. Section 8.2 will describe the baseline system for extracting the log-likelihood scored for each phoneme in the corpus and will analyze its features and the influence of all the sources of variability in them. Section 8.3 will study different score normalization techniques: Traditional methods like Test Normalization (T-norm) or Zero normalization (Z-norm) and techniques based on phonetic N-Best lists. Moreover, Section 8.4 will present the results with a set of SD retrained models for the PV task. Finally, Section 8.5 will make a comparative analysis of the results in terms of EER of all the proposed techniques.

8.1 The Pronunciation Verification Task

The PV task presents some parallelisms to the speaker verification task as it can be seen in Figure 8.1. These similarities were seen as the basis to apply similar approaches than those of speaker verification into the PV problem.

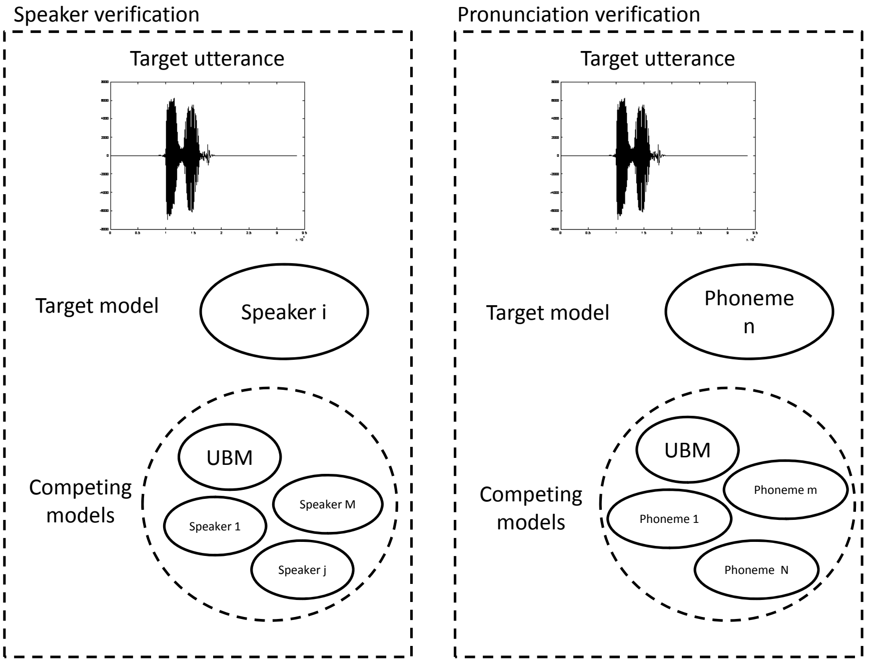


Figure 8.1: Resemblances between speaker verification and pronunciation verification

In speaker verification, the trial utterance (or a segment of speech obtained from speaker segmentation) is hypothesized to have been uttered by an speaker i from a group of M speakers. The most popular approaches for speaker verification use GMM-based models to calculate the likelihood of the input signal to the speaker model, with these speaker models having been previously trained with a sufficient number of utterances by each speaker. This score is then compared to the score obtained by a Universal Background Model (UBM) trained from a sufficiently large number of speakers. Impostor models, trained specifically for the rest of speakers $j = 1...N$ with $j \neq i$ are also used in different techniques to improve the performance of the system.

In phoneme verification, the trial utterance is hypothesized to be a sequence of phonemes and phoneme segmentation is used to detect phoneme boundaries in the utterance. Once a certain segment has been separated and hypothesized to have been generated by the phoneme n from the group of N phonemes, the score obtained by this model is compared to a UBM trained by all phonemes, to a garbage model of the phoneme n or to all the rest of phonemes ($m = 1...N$ with $m \neq n$) to detect whether that segment really corresponds to the proposed phoneme. As phoneme models have to deal with coarticulation and temporal variances in the signal, which are usually neglected in the speaker model in text independent speaker verification, HMM are used (as in ASR) instead of GMM.

8.1.1 Sources of variability

The basis of all speaker verification systems is to separate the different speech features that each possible speaker has as part of the speaker variability existing in a group of different speakers. The main deal of research up to date in speaker verification has been the elimination of those sources of variability in the speech signal that can hide these speaker features, mainly channel and session variability. In the same way, PV will suffer from the problem of those sources of variability in the speech signal different from the phoneme variability. As it is shown graphically in Figure 8.2, speaker variability and channel variability will have the same effect on the performance of PV than channel and session variability in speaker verification.

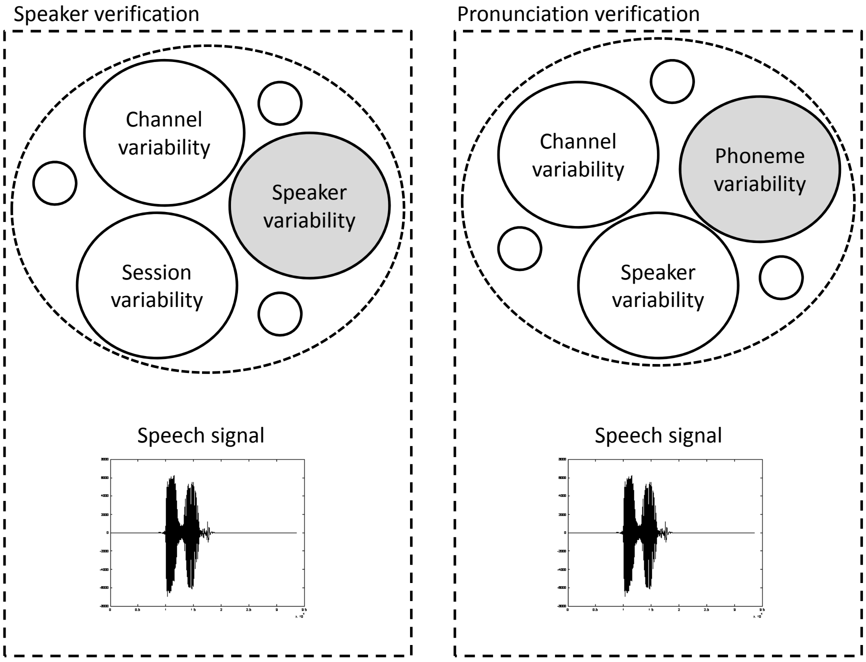


Figure 8.2: Sources of variability in speaker verification and phoneme verification

Different techniques have been proposed in the literature to avoid channel variability in speaker verification tasks:

Score normalization techniques are used to reduce the effect of the channel in the speaker's score. Cohort normalization, T-norm and Z-norm are the most usual of them [Auckenthaler et al., 2000]. While T-norm normalizes to the scores of the competing speakers in the same test signal, erasing hence the effect of channel in the normalized score, Z-norm normalizes to the scores of the target speaker model in different impostor signals, requiring a previous training phase.

Retraining of models [Kenny et al., 2003] consists in the training of several speaker models for all the possible channel conditions. This technique is effective as it separates different channel effects in different models. The speaker test signal is, then, evaluated with the speaker model that matches the channel conditions of the input signal. Drawbacks of this method are that sufficient data are necessary from the speaker in all channel conditions to train separate models and that it is necessary to know a priori the channel condition of the test signal or to have a robust estimator of it, because the use of channel mismatched models can have a dramatic effect in the performance of the system.

More complicated techniques like *Joint Factor Analysis* [Kenny et al., 2006] aim to separate speaker and channel variability by training separately speaker models in all channel conditions (eigenvoices) and channel models with speech from all speakers (eigenchannels). In the test signal, the effects of the speaker and the channel are separated via a factor analysis, allowing to evaluate the speaker model on solely the speaker space.

Once it has been argued how the PV task shares similar conditions with the speaker verification task, it was straightforward to think that some of the techniques applied in speaker verification could be used by PV. More exactly, score normalization and model retraining were studied and analyzed in different situations and conditions for the proposed task.

8.2 Experimental Framework and Score Characterization

8.2.1 Score extraction system

The baseline score extraction system for the experimental framework followed the diagram in Figure 8.3. A Viterbi based forced alignment set in an automated way the estimated phoneme boundaries within the input utterance. The log-likelihood score obtained by the model of the target baseform phoneme for each segment was calculated and finally, compared with a threshold to make a decision of accepting or rejecting the pronunciation as correct or incorrect and provide the evaluation decision.

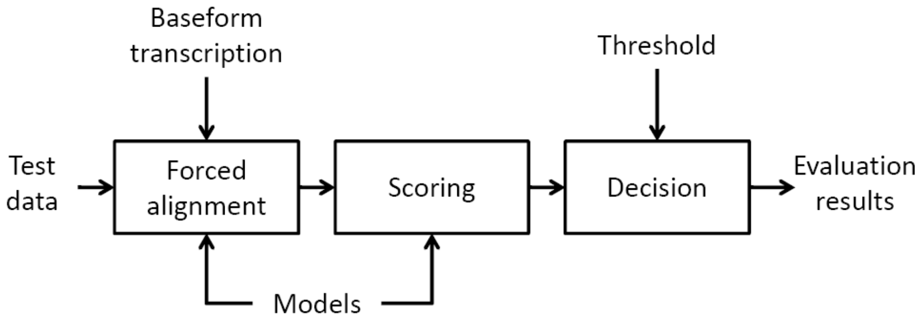


Figure 8.3: Baseline score extraction system for PV

The log-likelihood score was obtained in Equation 8.1 as the logarithm value of the likelihood probability of the segment of speech x being generated by the model of phoneme i (λ_i) averaged by the total number of frames (N_i) assigned to the given speech segment.

$$LL(i) = \frac{log(P(x|\lambda_i))}{N_i} \tag{8.1}$$

The likelihood probability of the speech segment and the model was calculated for the N_{s_i} frames of the segment of speech x as the probability of each frame (t_n with $n = 1...N_{s_i}$) over the GMM (g with $g = 1...G_s$) of G_s Gaussians that defined the current state (s with $s = 1...S$) of the phoneme HMM of S states, as in Equation 8.2. The unit and the state within the unit were obtained in the forced alignment phase in Figure 8.3.

$$P(x|\lambda_i) = \sum_{s=1}^{S}(\sum_{n=1}^{N_{s_i}}(\sum_{g=1}^{G_s}(p(g)p(t_n|g))) \tag{8.2}$$

To restrict the wide dynamic range that the log-likelihood scores could present (in the interval [$-\infty$,$+\infty$]) and have a more controllable range for the setting of the threshold, the sigmoid function in Equation 8.3 was used to compress the results to the interval [-1,$+1$].

$$LL_{Sigmoid}(i) = 2 * (\frac{1}{1 + e^{-LL(i)}}) - 1 \tag{8.3}$$

8.2.2 Experimental framework and performance measure

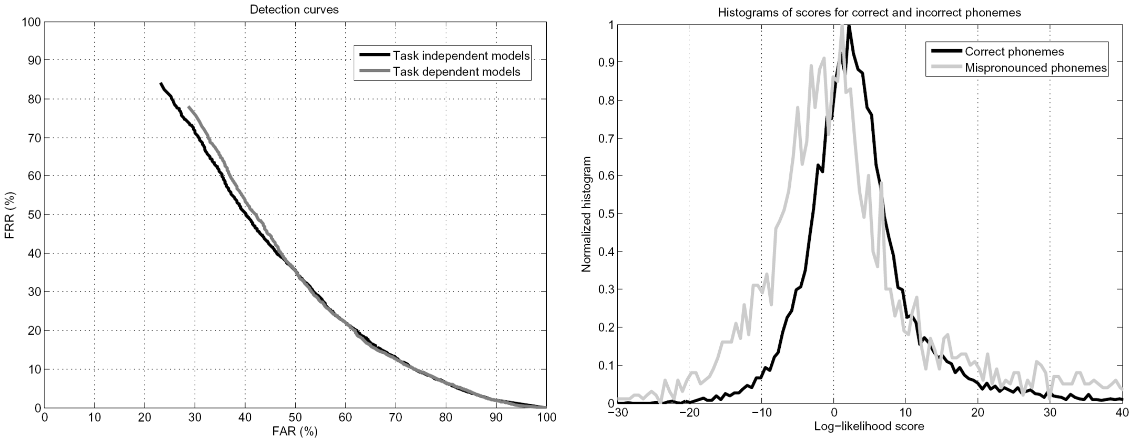
The experimental framework for the PV task was exactly the same that for the ASR experiments in Sections 4 and 7, in terms of the design of the acoustic models (number of units and states and Gaussians per unit), lexical models (transcription of the words in the vocabulary to acoustic units) and feature extraction system (dimensionality of the feature vectors, etc). Sub-phonetic context dependent models were used in the Viterbi forced alignment, as the results in ASR showed their good performance to detect segments of speech. Phoneme models were used instead in the scoring system because of their generalization of coarticulation boundaries, which provided certain independence with the scores of the neighboring phonemes.

The performance of all the proposed systems was made in terms of the detection curves and EER values. The calculated log-likelihood scores went through the threshold-based detection system and all possible thresholds in the sigmoid function interval were evaluated. For each threshold value the FAR and FRR were obtained, with FAR the rate of phonemes labeled as mispronounced (substituted or deleted) that were accepted as correct by the decision system and FRR the rate of phoneme labeled as correct that were not accepted by the decision system.

The detection curves have been widely used in detection problems and plot both values (FAR and FRR) for all possible thresholds in a continuous line; while the EER has been widely accepted as a comparative measure of performance of these systems indicating the value of FAR and FRR for the threshold in which $FAR = FRR$.

8.2.3 Baseline results and analysis of scores

All the 3,192 signals in the corpus of disordered speech were fed to the baseline PV system proposed previously, resulting in the evaluation of 16,351 phonemes (13,472 correct phonemes and 2,880 mispronounced phonemes according to the human experts). The TI and TD models used in the ASR experiments in Chapter 4 were tested also in this task to determine whether there could be any influence of the task and domain in the results. Results are showed in Figure 8.4(a) for both models, with an EER of 43.84% for the TI models and 44.80% for the TD models.



(a) Detection curves for TI and TD models

(b) Score histograms for correct and mispronounced phonemes

Figure 8.4: Detection curves and score histograms for the baseline system

The results showed an extremely poor performance of the proposed scores, close to the worst possible scenario of 50% with very badly behaved detection curves, showing the extreme influence of the different sources of variability in the signal. This effect was corroborated by the normalized histograms of the log-likelihood scores before the sigmoid function for correct and mispronounced phonemes in the TD case shown in Figure 8.4(b). Both histograms presented very little separability and a deep confusion between both types of phonemes appeared at the score level.

8.3 Score Normalization

Once seen the effect of the different sources of variability in the baseline results of the PV task, the study of possible techniques to avoid this effect started with the use of different normalization techniques: Starting from traditional techniques in speaker verification like T-norm and Z-norm, normalization techniques adapted to the special characteristics of the PV task were proposed [Saz et al., 2009e], which in the end will finish resembling traditional approaches in this task. The normalization block is introduced in the detection scheme explained in previous Section as seen in Figure 8.5, after the score calculation and prior to the sigmoid function and threshold decision.

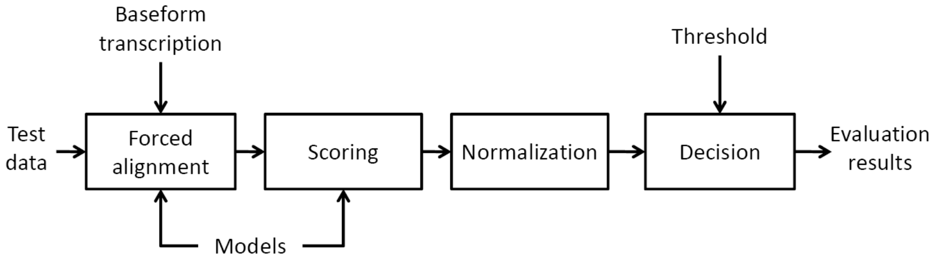


Figure 8.5: Score extraction system for PV with score normalization

8.3.1 T-norm and Z-norm evaluation

T-norm [Auckenthaler et al., 2000] is a well known technique for restraining variability in speaker verification systems. It is based on the Gaussian properties of the scores achieved by different speakers over a same speech segment. The basis of T-norm for a given speech segment, whose target speaker is p and the log-likelihood score achieved by that speaker is $LL(p)$, is to calculate all the log-likelihood scores on the utterance for a large number of impostor speakers and their mean ($\mu_{utterance}$) and standard deviation values ($\sigma_{utterance}$) for, posteriorly, perform a Gaussian normalization like in Equation 8.4.

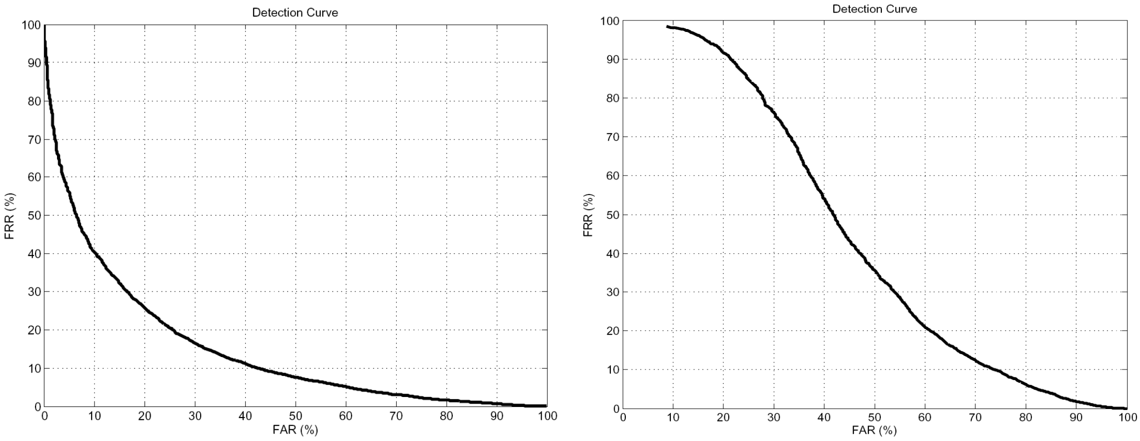
$$LL_{T-norm}(p) = \frac{LL(p) - \mu_{utterance}}{\sigma_{utterance}} \tag{8.4}$$

For the proposed task of PV, some aspects of T-norm had to be changed to adequate it to the task. In PV, the Gaussian statistics ($\mu_{utterance}$ and $\sigma_{utterance}$) are calculated from the log-likelihood scores of all the competing phonemes on the phoneme segment. T-norm in speaker verification usually relies on a sufficiently large number of impostor speakers to correctly calculate μ and σ for a given speaker segment. In the new PV task, it was not possible to count on more competing phonemes apart from those of the language: 24 as the number of phoneme units was 25.

The evaluation of the T-norm in the PV task can be seen in terms of the detection curves in Figure 8.6(a), achieving a 22.80% of EER. It was seen how T-norm kept its good properties to erase sources of variability from the speaker verification task to the phoneme verification task, and achieved a major improvement over the poor non-normalized results in previous Section.

Z-norm proposes another alternative to the elimination of variability in these detection tasks. It performs a previous training phase where it calculates the statistics ($\mu_{speaker}$ and $\sigma_{speaker}$) of the log-likelihood scores of a given speaker model through a sufficiently large number of segments from other speakers. With these statistics, the Z-norm of the log-likelihood of a speech segment, hypothesized to belong to speaker p, is calculated as in Equation 8.5.

$$LL_{Z-norm}(p) = \frac{LL(p) - \mu_{speaker}}{\sigma_{speaker}} \tag{8.5}$$



(a) T-norm

(b) Z-norm

Figure 8.6: Detection curve for T-norm and Z-norm

Again, in the PV task, the elements of the Z-norm had to be redesigned. In the training phase of the Z-norm parameters, each speech segment corresponding canonically to a certain phoneme was evaluated for all the competitors. Finally, each phoneme obtained its μ and σ as the Gaussian properties of all these values. Once again, there was a limit in this technique in the PV task, because each phoneme will only evaluate 24 possible different competitors, while in speaker verification it is aimed to face as many impostor segments as possible.

The experimental framework for the evaluation of Z-norm supposed the realization of a training phase to calculate μ and σ for all the 25 phonemes. This training was made over the 13,224 signals of speech from the unimpaired speakers. The evaluation of the Z-norm in the PV task can be seen in terms of the detection curves in Figure 8.6(b), achieving a 44.31% of EER. With these results in hand, Z-norm did not suppose any improvement over the system without score normalization, showing a major complication in the translation of this technique directly to PV.

8.3.2 N-best based normalization and Goodness Of Pronunciation

While T-norm was seen as successful in removing speaker variability from the disordered speech to better detect phonetic mispronunciations, there were some issues about it that did not make it fit totally within the specific PV task.

T-norm hypothesizes Gaussian statistics in the distribution of the scores obtained by a sufficiently large number of impostor speakers and, hence, models this distribution with the mean μ and standard deviation σ. These values are used for the Gaussian normalization in Equation 8.4, that measures the deviation between the target speaker score from the mean, in terms of the standard deviation. However, this Gaussian assumption could not be assured in the PV task because the number of impostors in the PV task was much more reduced that in speaker verification (only the 24 remaining phonemes), leading to a lack of independence between the scores of the different impostors, as phonemes with similar features (vowels, consonants, or phonemes sharing point or mode of articulation) will likely obtain closer scores.

With these assumptions, a normalization which made a better use of the scores achieved by the small number of competing phonemes was proposed. This type of normalizations, resembling cohort normalization [Rosenberg et al., 1992], required computing the log-likelihood scores for all the competing phonemes ($N-1$, as the total number of phonemes is N) in the same segment of speech that the canonical unit i ($LL(i)$): $LL(n)$ for $n = 1...N$ with $n \neq i$.

Once all the scores of the competing units were calculated, several techniques could be applied to use them in normalization of the score of the baseform phoneme. Two of them were studied in this work: A GOP-like normalization and a 1-best normalization.

The GOP-like normalization applied the same framework as proposed for the calculation of GOP [Witt and Young, 1997, Witt and Young, 2000], which is a very well known technique for phoneme-level verification. In this case, the normalized log-likelihood was computed as in Equation 8.6, subtracting the log value of the sum of the scores for all the competing phonemes. The only difference that this normalization approach had with GOP was that the averaging by the number of frames is computed inside the log-likelihood instead of for the whole posterior.

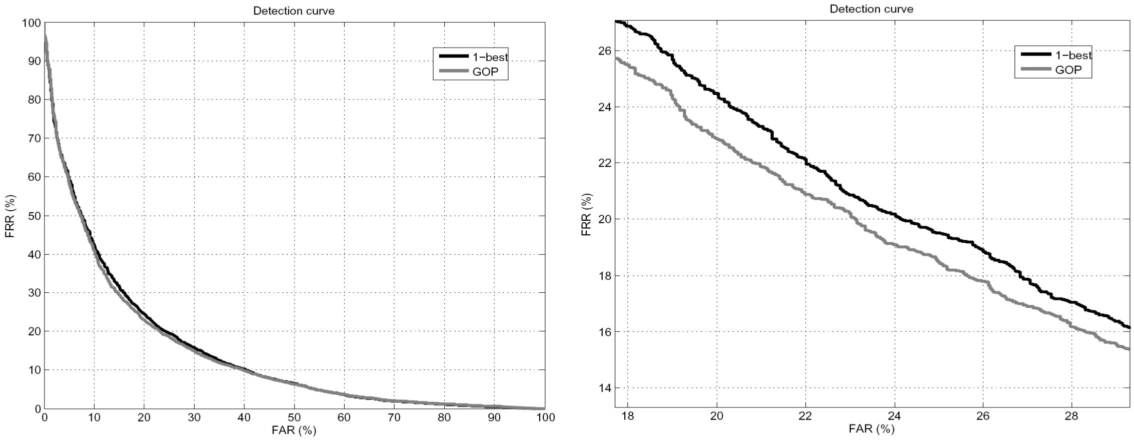
$$LL_{GOP_norm}(p) = LL(p) - LL(\frac{1}{N}\sum_{n=1}^{N}(P(s|\lambda_n))) \tag{8.6}$$

Traditional approaches to GOP approximate the calculation of the sum of all the scores of the competing phoneme to the score of the phoneme with the maximum score [Witt and Young, 2000]. This approximation, although popular, is quite gross as it is expected that several phonemes might have comparable values of scores, thus, limiting the mathematical application of the approximation.

However, with the proposed normalization approach, any approximation to the second factor in Equation 8.6 is plausible in a same way to the already mentioned cohort normalization. Hence, it was also evaluated a 1-best normalization where the log-score of the best competing phoneme (i.e. l with $LL(l)$) was directly normalizing the log-score of the baseform phoneme as in Equation 8.7.

$$LL_{1-best}(p) = LL(p) - LL(l) \tag{8.7}$$

This approximation limited the number of calculations required for the score normalization, as some phonemes could be pruned before computing their final log-score value and, furthermore, it provided directly an alternative phoneme as the best phoneme in the N-best list of alternate phonemes.



(a) Full detection curve

(b) Detail of the EER zone

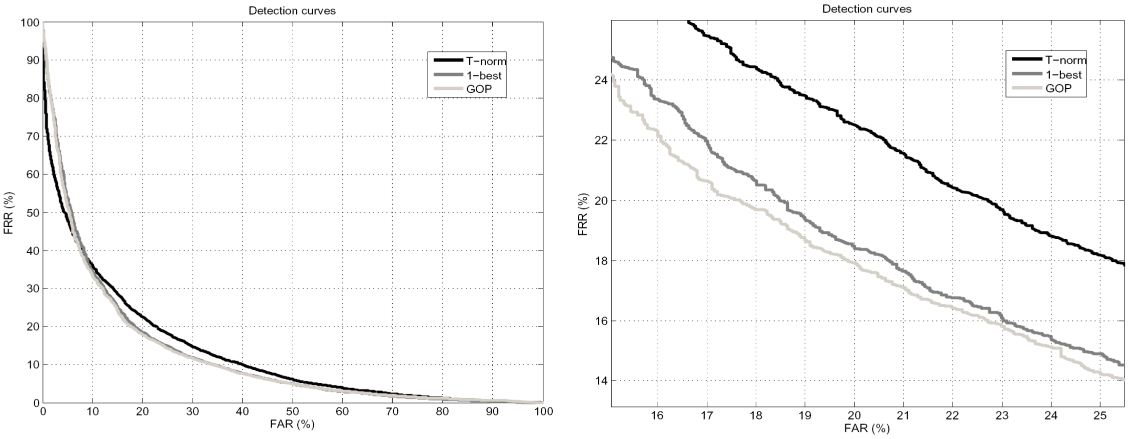
Figure 8.7: Detection curve for N-best based normalizations

Both normalization proposals achieved the detection curves shown in Figure 8.7, with a 22.01% of EER for the 1-best based normalization and 21.44% for the GOP-like normalization. Both techniques improved T-norm and showed fine properties in their detection curves and EER. Furthermore, several proposals could raise from the N-best list as different combinations of the scores from the competing phonemes could be used according to different phonetic properties of the unit for evaluation.

8.4 Speaker Adaptation

As explained in Section 8.1, retraining of the speaker models to different channel conditions is also an effective way of erasing channel variability in the speaker verification task. In a similar way, the retraining of speaker adapted models was tested in PV to improve the results of the previous Section.

These adaptation experiments were performed following the same framework that the speaker adaptation experiments done for ASR in Chapter 7, trying to investigate the same effects in amount and phonetic accuracy of the data. Furthermore, the normalization techniques studied in the previous Section were tested to study their performance with SD models.



(a) Full detection curve

(b) Detail of the EER zone

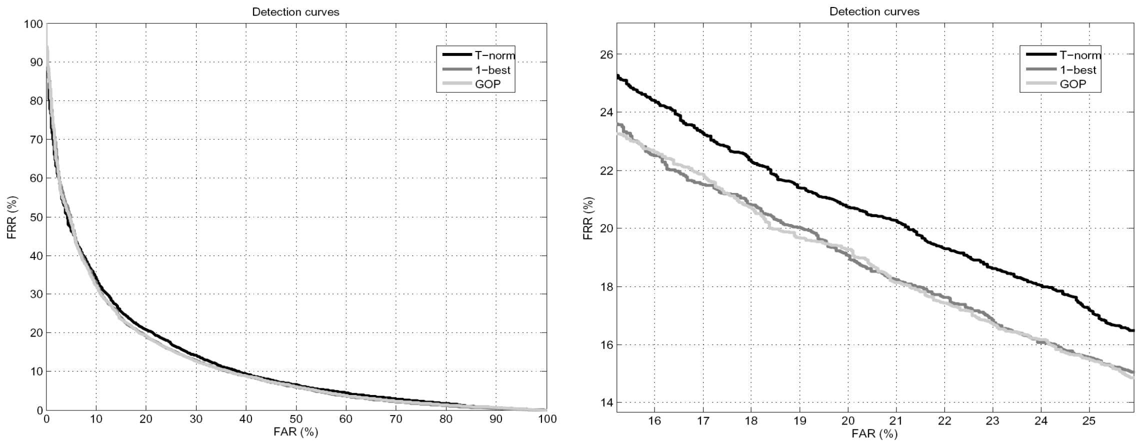
Figure 8.8: Detection curves for speaker adapted models to the baseline transcription

The first adaptation strategy carried out was to retrain the models according to the baseline transcription, using the same models than in the first approach in Section 7.1. The detection curves with the 3 normalization techniques can be seen in Figure 8.8. T-norm achieved a 21.29% of EER, while the 1-best based normalization achieved 19.17% EER and the GOP-like normalization achieved a 18.87% EER.

Second approach was to use the outcome of the APD system presented in Chapter 4 to obtain a new set of transcriptions as in the acoustic adaptation of models for ASR in Section 7.1. The detection curves for the 3 normalization techniques can be seen in Figure 8.9. T-norm achieved a 20.46% of EER, while the 1-best based normalization achieved 19.55% EER and the GOP normalization a 19.48% EER.

Posteriorly, the SD models trained following the outcome of the experts' labeling, explained in Section 7.1, were also used for this same PV task. The detection curves in Figure 8.10 showed a 18.16% EER for the T-norm; and a 15.63% EER for the 1-best based normalization and 15.71% for the GOP-like normalization.

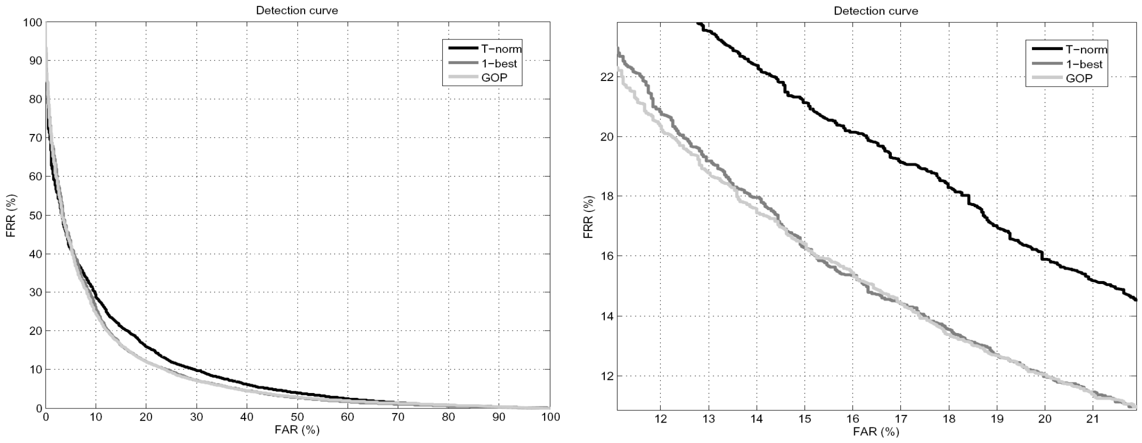
In all cases, the results outperformed the EER values achieved with the SI models in the previous Section. As it will be seen in the discussion of results, the PV task was seen as very sensitive to the accuracy of the pronunciation in the data used for the retraining of models.



(a) Full detection curve

(b) Detail of the EER zone

Figure 8.9: Detection curves for speaker adapted models to the APD transcription



(a) Full detection curve

(b) Detail of the EER zone

Figure 8.10: Detection curves for speaker adapted models to the experts' labels

8.5 Comparison of Techniques

Once presented the results for all the score normalization techniques proposed altogether with all the possibilities for speaker adaptation of the acoustic models, this Section gathers all these results for a better interpretation of all of them. Table 8.1 presents all the values of EER from the Chapter.

As it could be seen, GOP normalization and the proposed 1-best normalization showed an improvement of up to 14% over T-norm, with both approaches showing no significant different in their results. Proposals on how to use more phonemes in the N-best list according to some rules might achieve better results than these two proposals, as GOP might be using information from irrelevant phonemes and the 1-best case might be limited for neglecting a large number of competitors.

An interesting result from this experiment was to observe how the SD model trained with only that data which was considered lexically accurate by the labelers outperformed the SD model trained from all the data considering the baseform transcription by a 14-18% for all normalization techniques. These results were compared to the performance of both models in ASR in Section

7.1, where the model trained with all the data outperformed the model trained with only correct data. Models trained for PV require more precision in the separation of the phoneme variability between correct and incorrect data. In this case, the use of mispronunciations in the training phase was seriously producing a loss in the quality of these models.

Table 8.1: EER values for the normalization techniques and speaker adaptation

	Acoustic adaptation			
Normalization	No adaptation	Baseform	Human labels	APD
T-norm	22.80%	21.29%	18.16%	20.46%
GOP	21.44%	18.87%	15.71%	19.48%
1-best	22.01%	19.17%	15.63%	19.55%

8.5.1 2-pass system

Once seen that the use of mispronounced phoneme units for the retraining of the models in PV degraded the performance of the system, compared to the "Wizard of Oz" system which only used for retraining the units considered as correct by the human experts, it was decided to study the possibility of developing a system that used this feature without the use of human help.

A preliminary unsupervised system for achieving the best possible performance was, hence, studied by means of a 2-pass system. The scheme of the system, in Figure 8.11, used as seed the TD models, and performed a PV step with these models and the system with the 1-best based normalization. The decision made over the phonemes in the first step was used in the re-training phase where only the data considered as correct by the PV was fed to the MAP algorithm. These new models were used in the second pass of the PV system to obtain the definitive scores and make the final decision of correctness or not of each phoneme for evaluation.

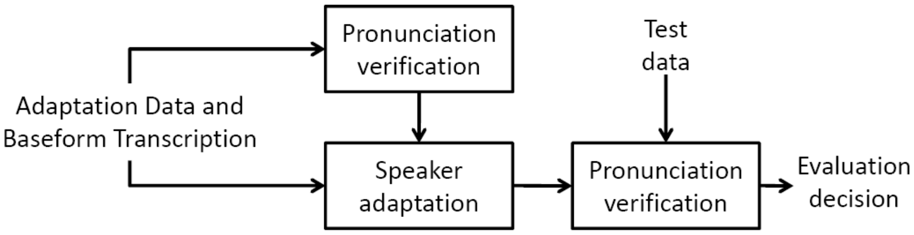
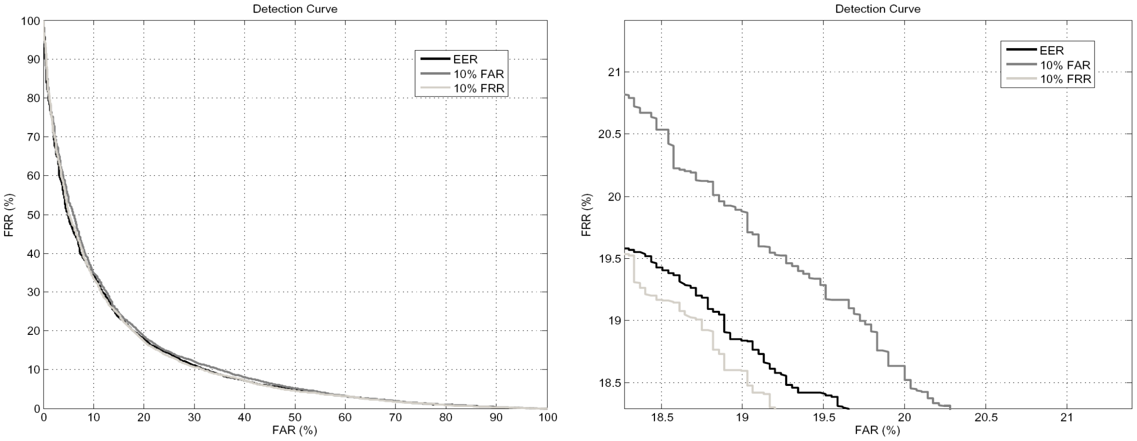

Pronunciation verification
Test data
Adaptation Data and Baseform Transcription
Speaker adaptation
Pronunciation verification
Evaluation decision

Figure 8.11: Scheme of the 2-pass unsupervised PV system

This 2-pass system was very sensitive to the precise tuning in which the first PV system was configured; and for that, three operating points where chosen and studied for this system: The first one was the EER point (threshold in -0.2), the second one is the point with only 10% of false acceptance (threshold at 0.37 and approximately 40% of false rejections) and the third one was the point with 10% of false rejections (threshold at -0.64 and approximately 40% of false acceptances). Working points situated in the extremes of the detection curve (thresholds above 0.5 or below -0.7) would lead to the situations of no adaptation or adaptation with all the data (correct and mispronounced phonemes) that are already known in columns 1 and 2 of Table 8.1.

The three detection curves achieved by the three working points are plotted in Figure 8.12 (full curve in 8.12(a) and zoomed around the EER area in 8.12(b)). The EER for the three working points are 18.91%, 19.38% and 18.82% for the EER, 10% false acceptance and 10% false rejection respectively.



(a) Full detection curve (b) Detail of the EER zone

Figure 8.12: Detection curves for 2-pass PV system

It was seen how the best operation points for the first PV phase in the 2-pass system were those who tried to use more data, achieving a 14-15% of improvement over the SI system and a 1.5% of improvement over the SD 1-pass system that uses all data for re-training. This gain had no significance to outperform the system that used all units without considering whether they were correct or not, and it was still far from the "Wizard of Oz" system that achieved a 15.63% EER by retraining only with the phonemes considered correct by the human labelers. The reasons for this was that the system was working in the trade-off point between two requirements in the adaptation. In one hand, it required the most possible amount of data, as it was seen how it was sensitive to this subject, while requiring the most accurate possible data.

Furthermore, phonemes incorrectly verified in the first stage would most likely be those who were closer to the border between the spaces of the correct and incorrect phonemes; becoming the most pernicious in the second phase, due to the higher confusion among them. Hence, discriminative training or large margin techniques might show a usefulness to provide an enhanced unsupervised retraining of models for the PV task.

Chapter 9

Unsupervised On-line Systems for ASR of the Speech Handicapped

I made him just and right,
Sufficient to have stood, though free to fall

-John Milton, *Paradise Lost*

Chapter 7 has introduced supervised techniques for acoustic-lexical adaptation. They required to know the word prompted to the user and/or the outcome of the labeling process to know which parts of the input signal matched the canonical transcription of the word. On the other hand, a framework for the detection of phonetic mispronunciations has been presented in Chapter 8 with promising results.

The base of unsupervised speaker adaptation is to feed the outcome of an ASR system as transcription to the adaptation algorithm with the input speech signal. The use of a confidence measure is requested to avoid feeding incorrectly recognized utterances or parts of utterances that may corrupt the output model.

Once unsupervised adaptation has been studied, on-line personalization will arise as the next step in the study of frameworks for providing adaptive ASR technology for special users. In this case, constant retraining is performed as the speaker continues using the ASR system, instead of limiting the adaptation to the initial frames of interaction. With these on-line approaches, re-adaptation can be achieved as the speakers change the acoustic-lexical properties of their speech, due to improvements in their language or to a worsening in their diagnosis.

The Chapter is organized as follows: Section 9.1 will introduce the problem of unsupervised adaptation and the baseline results when no confidence measuring is performed over the transcriptions fed to the adaptation algorithm. Section 9.2 will introduce all the strategies taken for providing unsupervised adaptation and Section 9.3 will present and try to solve the problems associated to on-line unsupervised adaptation.

9.1 Proposal and Baseline of Unsupervised Adaptation

As mentioned previously, the objective of an unsupervised adaptive system is to train new SD models with a set of input signals whose transcription is initially unknown. The reason for the evaluation of these unsupervised frameworks arose in the Introduction in Chapter 1 due to the impossibility to reliably count on enrollment phases when working with young handicapped users.

With these limitations, the proposed system had to adapt itself automatically using the same data of the ASR interaction.

To provide of comparability between all the supervised adaptation frameworks presented in Chapter 7 and the results in this Chapter, the strategies for adaptation were made similarly to those of the previous Chapters. Hence, 3 of the 4 sessions were considered for re-training in 4 leave-one-out experiments, where the achieved models were tested over the remaining session. The only information taken from the training data were the input speech frames as no knowledge on the uttered word or the labeled mispronunciations was used.

The baseline for unsupervised adaptation was the simple framework in Figure 9.1. The training signals were fed to the initial TD ASR system (the same that in Chapter 4), which provided a word transcription of each signal. All transcriptions were used to re-train the new model which was used with the remaining test signals to achieve the final results.

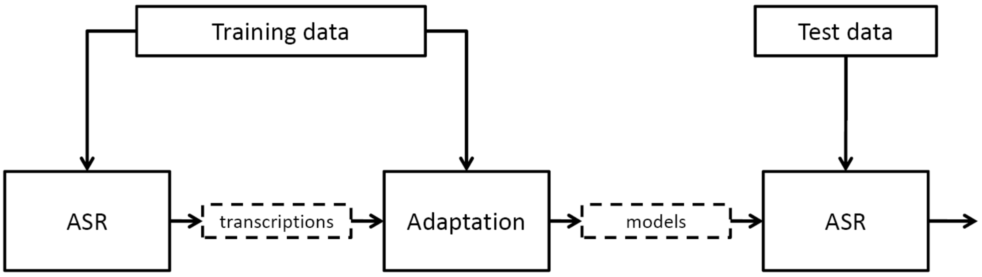


Figure 9.1: Simple unsupervised adaptation framework

The results for the test signals in the proposed framework can be seen in Table 9.1 for the three possible acoustic units (words, phones and sub-phones). The first ASR decoding phase with the training signals was performed with the same type of units that the final SD ASR experiments.

Table 9.1: WER in the simple unsupervised adaptation framework

Speaker	word	phone	sub-phone	Speaker	word	phone	sub-phone
*Spk*01	4.82%	4.82%	5.26%	*Spk*02	15.79%	14.91%	18.42%
*Spk*03	3.95%	3.07%	2.19%	*Spk*04	2.63%	2.19%	2.19%
*Spk*05	63.16%	56.14%	51.75%	*Spk*06	1.75%	2.63%	1.75%
*Spk*07	20.18%	21.93%	19.30%	*Spk*08	42.11%	42.11%	43.42%
*Spk*09	16.23%	14.04%	14.91%	*Spk*10	19.74%	29.82%	26.32%
*Spk*11	5.70%	7.02%	7.89%	*Spk*12	53.51%	58.33%	56.14%
*Spk*13	78.95%	73.25%	79.82%	*Spk*14	13.60%	9.21%	10.53%
AVG	24.44%	24.25%	24.28%				

The average WERs in Table 9.1 were around 24% for all units, which supposed a relative improvement with respect to the TD baseline results (Table 4.5 in Section 4.3) of 14-20%. The results showed the special difficulty of the unsupervised task in the comparison with the 12-15% WERs achieved in Section 7.1 when the prompted word was known (supervised case).

9.1.1 Lexical unsupervised adaptation

Lexical adaptation was also evaluated in an unsupervised framework. The steps that were taken for this were as shown in Figure 9.2, where the adaptation data was fed at the same time to the SI-TD ASR system to the SI-TD APD. ASR obtained the proposed word transcription for each utterance, while APD provided the transcription of the utterance in phone/sub-phone units. Posteriorly, the decoded word and the transcription were matched and added to the new SD lexicon for each speaker.

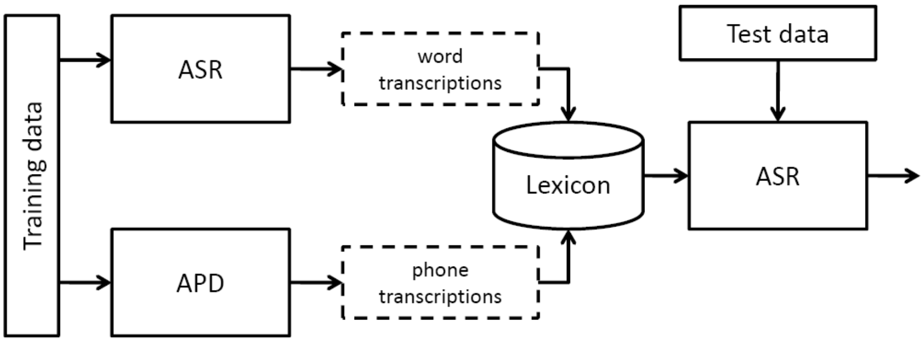


Figure 9.2: Simple unsupervised adaptation framework

The leave-one-out experiments were replicated here with 3 sessions for adaptation and 1 session for testing and obtaining the final results.

The results achieved by all the speakers are shown in Table 9.2. A noticeable and significant decrease of performance was observed with respect to the baseline TD results (from 28-31% to 31-34% WER). However, this increase in WER is fully justified by the direct use of the output of the ASR system into the new lexicons. When a recognition mistake is done in the preliminary recognition phase, this error was propagating to the evaluation phase through its inclusion in the lexicon.

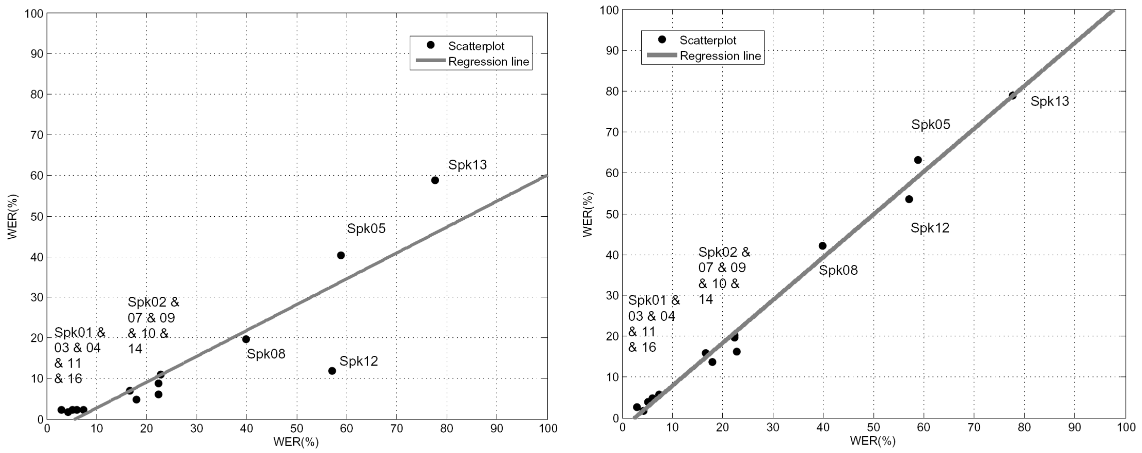
Table 9.2: WER in the simple unsupervised lexical adaptation framework

Speaker	word	phone	sub-phone	Speaker	word	phone	sub-phone
*Spk*01	-	13.60%	11.43%	*Spk*02	-	27.63%	21.49%
*Spk*03	-	20.18%	9.21%	*Spk*04	-	4.39%	3.07%
*Spk*05	-	64.91%	63.60%	*Spk*06	-	10.96%	5.70%
*Spk*07	-	33.33%	32.89%	*Spk*08	-	56.14%	46.93%
*Spk*09	-	26.32%	21.93%	*Spk*10	-	38.16%	38.16%
*Spk*11	-	14.91%	9.21%	*Spk*12	-	73.68%	70.61%
*Spk*13	-	75.88%	83.33%	*Spk*14	-	25.44%	19.74%
AVG	-	34.68%	31.24%				

Hence, for lexical adaptation, even much more than for acoustic adaptation; the use of a confidence measure was requested to limit the influence of ASR inaccuracies into the performance of the system. All this process was performed and will be shown in Section 9.2.

9.1.2 Effect on the unsupervised results of the initial ASR mistakes

A study was made to understand how different levels of misrecognition in the first ASR decoding phase affected the final performance of the simple proposed unsupervised system. Figure 9.3 presents the scatterplots that related the WER for all the speakers in the word-unit supervised results in Section 4.3 and the WER in the word-unit unsupervised results recently shown, both against the WER of the word-unit baseline TD results.



(a) Supervised system

(b) Unsupervised system

Figure 9.3: Relations between WER in the first and second phase

Figure 9.3(a) represents the scatterplot of WERs in the supervised system vs the WERs of the SI-TD system, with their regression line ($y = 0.64x - 3.6868$). Although the regression was not significantly linear, as r was only 0.90, it was possible to see how they related and their trends.

On the contrary, Figure 9.3(b) shows the same scatterplot in the unsupervised case. The regression line in this case was $y = 1.05x - 2.71$, with r reaching 0.99, showing the bad characteristics of this system, because the slope of the line was above 1, which meant that in the upper zone of the regression line, the WER of the unsupervised SD models were greater than the WER of the SI-TD baseline system. And this was actually happening for the speakers with highest WER ($Spk05$, $Spk12$ and $Spk13$), whose results got degraded with the unsupervised adaptation.

The ideal situation would be one in which the WER after the adaptation was constant, independently of the personal characteristics of the speakers and, hence, independent too of the WER of the speaker with SI models. In the unsupervised case, the dependency between SI and SD recognition experiments is greater, as mistakes in the initial ASR propagate through the retraining of models. The following research had to be oriented to limit these inaccuracies in the adaptation data which was fed to the unsupervised system, as the way to improve their performance.

9.2 Unsupervised Adaptation based on Confidence Measures

A more precise unsupervised adaptation system that could overcome the effect of misrecognitions in the initial decoding phase of the training data was proposed like in Figure 9.4. This system made use of a confidence measuring block to post-process the output transcription and decide whether the decoded word had enough confidence or not to be used for adaptation or to extract those parts of the utterance that could have higher confidence.

It was decided that, instead of rejecting utterances at the word level, the confidence measuring was made at the phoneme level, because this could provide an enhanced selection for two reasons:

- On one hand, parts of the input signal could be correctly associated to a certain phoneme, even if the utterance had been misrecognized.

- On the other hand, even in the case of an ASR success, it might happen that some phonemes were mispronounced by the speakers, as the human labelers had evaluated.

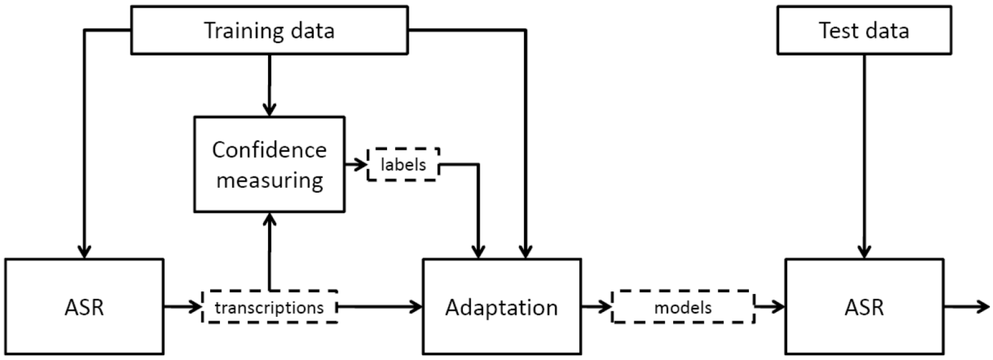


Figure 9.4: Unsupervised adaptation framework based on confidence measuring

The proposed system made use of the confidence measure system proposed in Chapter 8, which showed a fine performance in the task of PV. Although the task was different, because it was not deciding the suitability of the utterance to the sequence of input phonemes, the similarities made this proposal useful: In the new task, the system had to decide which parts of the utterance were really belonging to the sequence of phonemes of the word proposed by the initial ASR decoding system.

The results of this system are shown on Table 9.3. A very important feature of this system was the operating point in which the confidence measuring was working to accept or reject phonemes. Table 9.3 presents the results for a threshold of 0, which was seen to be a point close to the EER in Chapter 8 when using the 1-best score normalization technique.

Table 9.3: WER in the confidence measure-based unsupervised adaptation framework

Speaker	word	phone	sub-phone	Speaker	word	phone	sub-phone
*Spk*01	-	6.58%	4.39%	*Spk*02	-	18.86%	19.74%
*Spk*03	-	4.39%	3.07%	*Spk*04	-	3.07%	1.75%
*Spk*05	-	55.26%	53.95%	*Spk*06	-	3.07%	1.75%
*Spk*07	-	23.25%	22.81%	*Spk*08	-	43.86%	45.61%
*Spk*09	-	15.79%	17.11%	*Spk*10	-	33.33%	27.63%
*Spk*11	-	7.89%	7.46%	*Spk*12	-	65.35%	64.47%
*Spk*13	-	74.12%	78.95%	*Spk*14	-	11.84%	12.28%
AVG	-	26.19%	25.78%				

The results showed a loss of performance with respect to the unsupervised adaptation which used all the data for adaptation, without distinguishing correct from incorrect phonemes. The effect of data sparsity was present again at this point of the work, so it was decided to evaluate how different thresholds to accept or reject the phonemes in the recognized words affected the results.

Figure 9.5 presents graphically these results for phone and sub-phone units: The possible thresholds for the decision ranged from -1 to +1, as these were the limits of the sigmoid function used. A threshold of -1 supposed accepting all the phonemes, and the results are those of Table 9.1 in Section 9.1; while a threshold of +1 supposed rejecting all the phonemes, i.e. the baseline results with SI-TD models. The rest of the values produced a set of results that grew nearly monotonically from the lower confidence values to the greater. Whatever effect could have had the effect of using more accurate data for adaptation did seem to be overwhelmed by the fact that less data was available for adaptation, as the threshold was becoming more strict to reject phonemes, thus, reducing the amount of adaptation data.

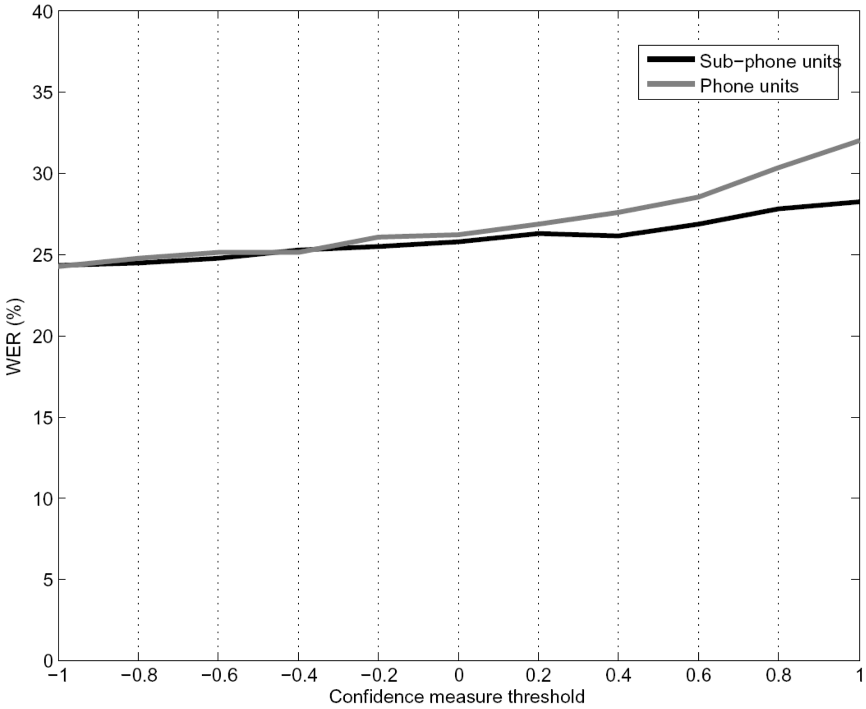


Figure 9.5: Unsupervised adaptation framework WER for different confidence thresholds values

Again, the effect of data sparsity was limiting the possibilities of providing a more accurate modeling of correct units when performing adaptation techniques to disordered speech. Although fast adaptation algorithms exist to avoid the problematic of sparse data, they are based on using different units to retrain a given unit, so they would be using more phonetically inaccurate data to the new HMM estimation, limiting the knowledge in how this data affected the final performance.

9.2.1 Lexical unsupervised adaptation based on confidence measures

A possibility to use confidence measure was furtherer evaluated via the framework shown in Figure 9.6. This framework proposed the rejection of possible recognition mistakes of the first phase via the phoneme-level PV system already used previously. The translation of the phone labels provided by the PV system to word labels followed a simple ruled based in the number of phonemes whose normalized score was lower than a threshold set to 0.

More precisely, when more than a 25% of the phonemes had been rejected considering that threshold, the word was rejected and it was not included with the corresponding APD transcription into the new SD lexicon.

With this technique, the number of lexical variants in the vocabulary was reduced to only those who were more accurate to the possible real pronunciation of the speaker. The results for all the speakers are presented in Table 9.4; a certain improvement was achieved over the lexical unsupervised adaptation presented in Section 9.1.1, but it still obtained worse performance than the baseline SI-TD results (3-7% worse relative performance).

Some other thresholds for the rejection of misrecognized words, according to their phone-level results in the confidence measure proposed, were evaluated, but they did not lead to any improvement over the shown results neither over the initial baseline. In any case, unsupervised lexical adaptation seemed possible for the proposed task.

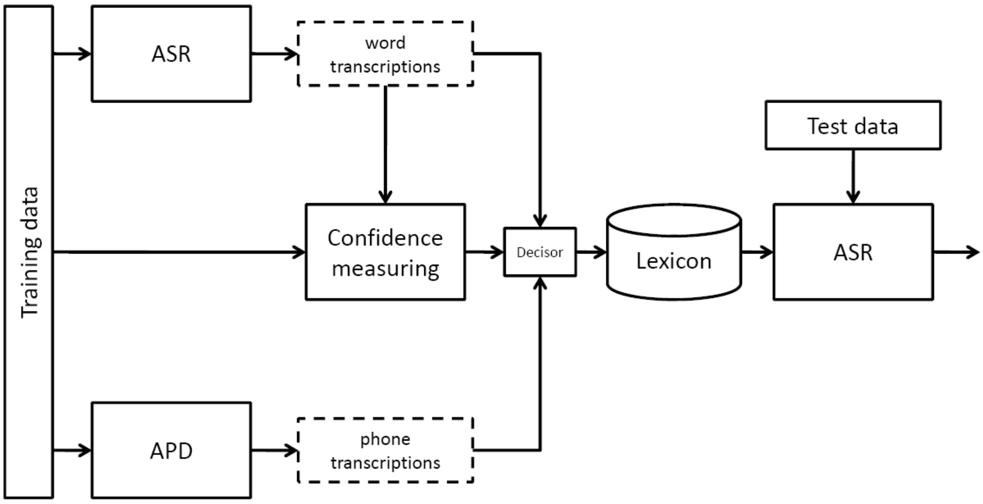


Figure 9.6: Unsupervised lexical adaptation framework with word rejection

However, it could be argued whether the proposal for confidence measuring-based unsupervised lexical adaptation was correct or not; and, even more, if it was really a plausible method in any case. The interest in lexical adaptation was to learn new pronunciations of a given word. Unfortunately, if a word got correctly recognized, but it corresponded to a different pronunciation, the confidence measure system would reject the word for considering it as incorrectly recognized, as there was no way for the system to distinguish incorrectly recognized words from correctly recognized words pronounced as lexical variants.

Table 9.4: WER in the unsupervised lexical adaptation framework with word rejection

Speaker	word	phone	sub-phone	Speaker	word	phone	sub-phone
*Spk*01	-	11.40%	10.53%	*Spk*02	-	23.25%	24.56%
*Spk*03	-	19.74%	8.77%	*Spk*04	-	3.95%	3.07%
*Spk*05	-	62.28%	58.33%	*Spk*06	-	9.65%	5.26%
*Spk*07	-	32.02%	27.63%	*Spk*08	-	46.05%	44.30%
*Spk*09	-	24.12%	22.81%	*Spk*10	-	36.40%	31.58%
*Spk*11	-	17.54%	11.84%	*Spk*12	-	74.56%	70.18%
*Spk*13	-	80.70%	80.70%	*Spk*14	-	21.49%	19.74%
AVG	-	33.08%	29.95%				

The problem of ambiguity in the recognition of this disordered speech arose, hence, stronger than ever; as it was seen than in most of the cases the output of the ASR system would be more accurate to the real pronunciation of the user than the way in which the speaker pronounced the word according to the baseform transcription of the prompted word.

This ambiguity was more problematic in lexical adaptation because, as mentioned before, there was no way for the automated system to discern between a correctly recognized word with a bad pronunciation from the user and a incorrectly recognized word whose pronunciation by the speaker was closer to this word than to the prompted one.

9.3 On-line Personalization of ASR Systems

After reviewing and evaluating the possibilities for supervised adaptation in Chapter 7 and unsupervised adaptation in previous Sections of this Chapter; this Section will make a proposal for an on-line unsupervised personalization system for ASR-based systems for the handicapped.

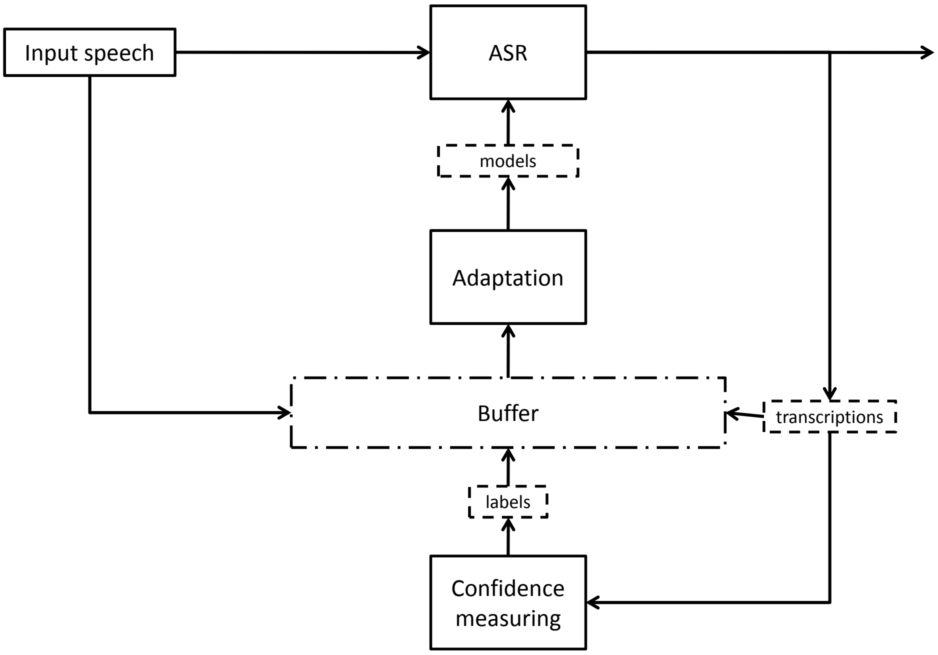


Figure 9.7: On-line personalization framework for ASR

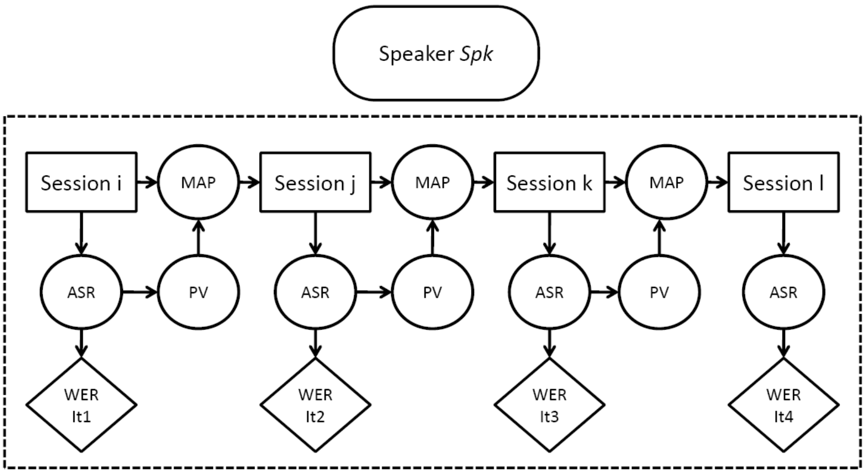


Figure 9.8: Experimental framework for on-line personalization

The proposed system followed the diagram in Figure 9.7. The user's input speech entered the ASR system to decode the oral utterance. A buffer stored each input speech signal, each decoded word provided by the ASR system and the phonetic labels provided by the confidence measuring system. When it was decided that enough data had been obtained for adaptation, the adaptation algorithm was fed on the contents of the buffer and applied the unsupervised adaptation shown in the previous section, obtaining a new SD model for ASR. After each adaptation, the buffer was emptied and refilled again with the new input data for further adaptation.

To simulate this situation with the available data, it was decided to perform little steps of adaptation with each session available from the speakers. Hence, the final process of work was as shown in Figure 9.8 where for each speaker *Spk*, Session *i* was initially recognized, evaluated with the PV system and models retrained for the recognition of Session *j* where the same process went on towards Session *k* and Session *l*. The total number of experiments was 24 per speaker, which was the number of ways in which 4 elements could be combined in a list of 4 without repetitions (4!). The final outputs were the averaged WERs after the first, second, third and fourth stages (iterations) of the process.

Experiments were run only over the sub-phone units, which had provided the best performance besides word units on the previous experiments. The results can be seen per speaker and each step of the iterative adaptive procedure in Table 9.5. The results after the first recognition step were the same as in the baseline TD results, as no adaptation had been carried out at this point. The subsequent phases of adaptation were providing an overall relative improvement of 5.43%, 8.58% and 11.56% over the SI results.

Table 9.5: WER of the on-line personalization system

Spk	It 1	It 2	It 3	It 4	Spk	It 1	It 2	It 3	It 4
*Spk*01	10.09%	5.99%	4.90%	4.24%	*Spk*02	20.18%	19.74%	19.08%	18.79%
*Spk*03	5.70%	4.82%	3.65%	3.07%	*Spk*04	3.51%	2.49%	2.19%	2.19%
*Spk*05	56.14%	54.53%	53.95%	53.22%	*Spk*06	3.07%	1.75%	1.90%	1.75%
*Spk*07	25.44%	24.85%	22.22%	19.52%	*Spk*08	43.86%	44.83%	45.98%	46.56%
*Spk*09	22.81%	18.57%	15.94%	14.40%	*Spk*10	32.46%	31.43%	30.56%	29.46%
*Spk*11	9.21%	7.89%	7.60%	6.29%	*Spk*12	63.60%	63.89%	60.67%	58.04%
*Spk*13	80.26%	79.82%	80.19%	81.07%	*Spk*14	18.42%	12.72%	12.06%	10.53%
AVG	28.20%	26.67%	25.78%	24.94%			-		

The WER in the final session (24.94%) improved the results in Table 9.3 which used the same framework for unsupervised adaptation with confidence measuring in an off-line format. In the previous case, 3 sessions were gathered to perform one single adaptation for the final session; on the contrary to the on-line situation, which performed small adaptation steps, and still got a better improvement. The on-line framework showed that constant retraining of models can be useful to provide bigger reductions in WER, despite the problem of data sparsity.

At this point, it was seen as important to make a deeper analysis of the improvement achieved separately for each speaker. If the proposed system was aimed to provide improvement in the recognition rates of an ASR system to be used by a disabled person, it was important to know which improvement could be expected depending on the characteristics of the target user.

For that reason, Table 9.6 shows the values of improvement achieved over the baseline results in each iteration of the on-line system for all speakers.

There were 8 speakers who achieved improvements with a high significance (over 20%) which were speakers *Spk*01, *Spk*03, *Spk*04, *Spk*06, *Spk*07, *Spk*09, *Spk*11 and *Spk*14. Other 4 speakers achieved improvements below 10%: *Spk*02, *Spk*05, *Spk*10 and *Spk*12; and, finally, two speakers (*Spk*08 and *Spk*13) had a negative performance of the personalization system. Again, the relationship with the WER of the initial recognition phase was clear, as the speakers with higher improvements were those whose baseline ASR results were around 20% of WER or lower. On the other hand, speakers with an initial recognition accuracy of 30 to 80% got minimum improvements within the personalization framework.

Table 9.6: Improvements of the on-line personalization system

Spk	It 1	It 2	It 3	It 4	Spk	It 1	It 2	It 3	It 4
*Spk*01	-	40.63%	51.44%	57.98%	*Spk*02	-	2.18%	5.45%	6.89%
*Spk*03	-	15.44%	35.96%	46.14%	*Spk*04	-	29.06%	37.61%	37.61%
*Spk*05	-	2.87%	3.90%	5.20%	*Spk*06	-	43.00%	38.11%	43.00%
*Spk*07	-	2.32%	12.66%	23.27%	*Spk*08	-	-2.21%	-4.82%	-6.16%
*Spk*09	-	18.59%	30.12%	36.87%	*Spk*10	-	3.17%	5.85%	9.24%
*Spk*11	-	14.33%	17.48%	31.70%	*Spk*12	-	-0.46%	4.61%	8.74%
*Spk*13	-	0.55%	0.09%	-1.01%	*Spk*14	-	30.94%	34.53%	42.84%
AVG	-	5.43%	8.58%	11.56%		-			

An open question that these experiments left was the performance with more data for continuing the iterative on-line procedure described in this Section, as it was expected that more new data would keep reducing the final WER for the impaired speakers, because the marginal improvements kept growing up from session to session.

From all this, it came clear that unsupervised techniques have to be applied preferentially to speakers whose disorders do not degrade largely the performance of the ASR systems, while users with heavier disorders should face a supervised enrollment phase, nevertheless the difficulty it may suppose for them.

Chapter 10

Development of Speech-based Applications for the Handicapped

> If someone makes a movie and no one sees it, does the movie exist or not?
>
> -Paul Auster, *Book of Illusions*

This Chapter presents all the real-world applications that have been developed during the realization of this thesis. One of the main motivations for this thesis was to transduce the theoretical and experimental results of the thesis to the development of applications that could improve the quality of life of a part of the handicapped community. These developments fructified in two types of tools: Development of speech-based tools for environment control for the handicapped and development of CALL tools with special focus on the handicapped community. These tools are the fruit of the work of the whole team at the VIVOLAB group in the I3A, other research groups at the University of Zaragoza and different institutions for the schooling and assistance of the handicapped. This Chapter will also be a recognition for all their effort.

The Chapter is organized as follows: Section 10.1 will introduce three different approaches for the control of different environment elements by people with impairments via speech and voice-based interfaces; while Section 10.2 will review the development of CASLT tools within the "Comunica" framework, including "PreLingua", "Vocaliza" and "Cuéntame". Finally, Section 10.3 will introduce the attempt in L2 learning tools with the development of "VocalizaL2" and an experience with this tool.

10.1 Speech-based Interfaces for the Handicapped

During the review of the stat-of-the-art in speech-based applications for the handicapped in Chapter 2, it was seen how the creation of interfaces adapted for handicapped people to operate with those elements whose physical difficulties would not allow them to use normally is a major field in this area of speech research. The Vocal Joystick [Harada et al., 2008] or the STARDUST and VIVOCA projects [Hawley et al., 2003, Creer et al., 2010] have been two successful examples of these developments, among others.

Three different elements for environment control were researched during the realization of this thesis. The first one was the introduction of a speech-based control system for an automatized wheelchair [Alcubierre et al., 2005]; the second one was the voice-based emulation of keyboard and mouse events by severely handicapped individuals [Saz et al., 2009a, Saz et al., 2009b]; and finally the inclusion of an ASR system in a virtual mouse element [González, 2009]. The framework of collaboration in which these actions were made, the technical characteristics of these control elements and the conclusions of the works are presented in the next sections.

10.1.1 Speech control of an automatized wheelchair

The development of sensor-based navigation systems for robots [Mínguez and Montano, 2005] has been applied during the latest years by the Robotics and Real Time Group of the I3A, led by Prof. Luis Montano and Dr. Luis Montesano and Dr. Javier Mínguez. One of the applications in which these systems have been used was a robotic wheelchair [Mínguez et al., 2006] able to detect obstacles with the help of a laser beam. This knowledge of the environment enabled the system to generate a safe route from the origin point to the destination point defined by the user.

A major line of research within the Robotics Group was the interest in possible interface systems for the control of this automatized wheelchair, developing different possibilities like touch screens, Brain-Computer Interface (BCI) [Iturriate et al., 2009] or speech-based systems, which are the subject of this Section.

The work carried out for the speech control of this wheelchair [Alcubierre, 2005] supposed the integration of the state-of-the-art ASR system used in the experimental results of this thesis into the control system of the wheelchair. For this purpose, there was to pay special attention to the Central Processing Unit (CPU) limitations imposed by the functioning of the laser detector, that blocked the CPU during most of its cycle to ensure the correct detection of the obstacles. The system was designed for the recognition of isolated commands via the rule-based grammar in Figure 10.1 in which an activation word ('Dusila') was used to indicate the ASR that the following commands were directed to the wheelchair. The commands within the grammar were related to the spatial directions in which the wheelchair could move ('arriba'-'up', 'abajo'-'down', 'izquierda'-'left', 'derecha'-'right', 'cerca'-'close', 'medio'-'medium' and 'lejos'-'far'); general commands like 'para'-'stop', 'anda'-'walk', 'inicia'-'start' and 'termina'-'finish' indicated the system when to start or stop the movement.

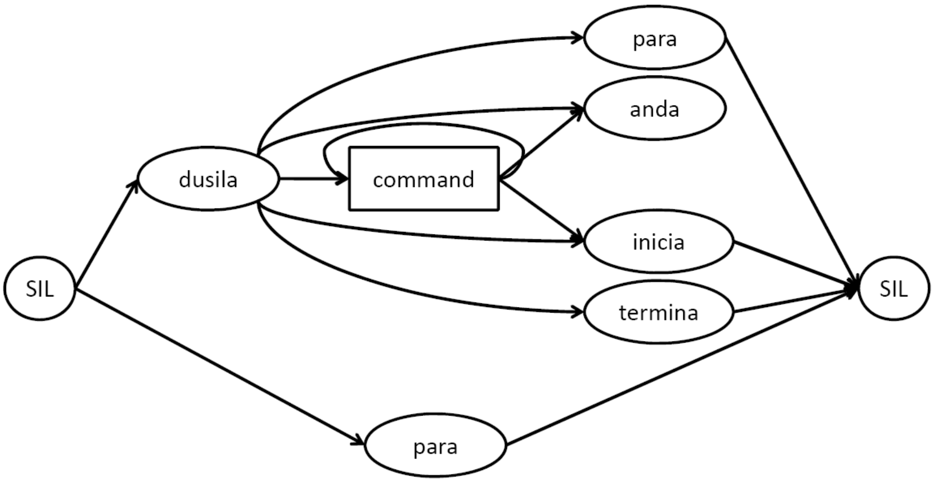
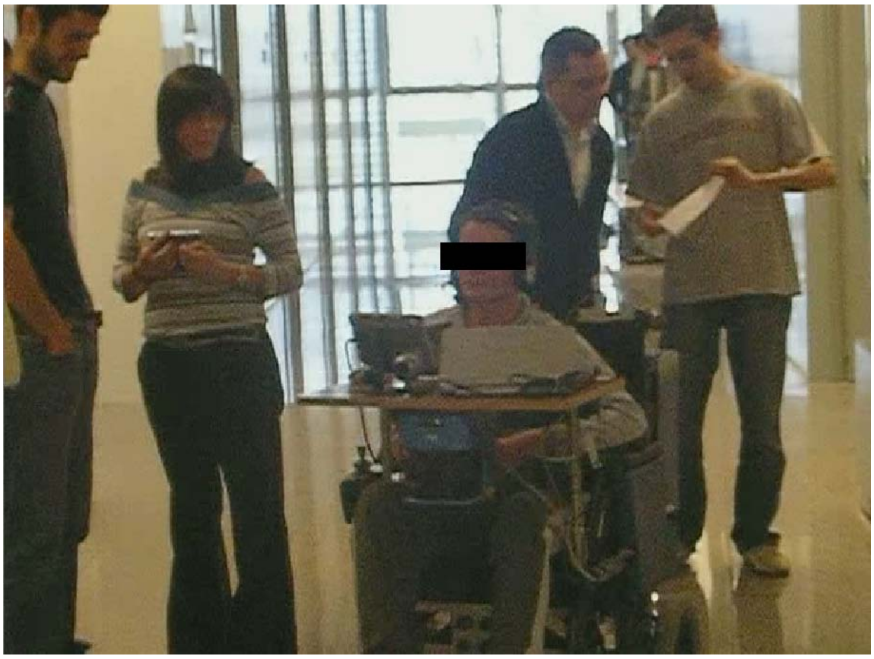


Figure 10.1: Grammar for the speech control of the automated wheelchair

The evaluation of this speech-based interface was carried out with a test subject, student with tetraplegia from the CPEE "Alborada" and was controlled by members of the CPEE "Alborada" and the Robotics group. A screenshot of a video shot during the tests (courtesy of Javier Mínguez) shows in Figure 10.2 the student in the wheelchair followed by Javier Mínguez and César Canalís from the CPEE "Alborada", while trying to make his way through the facilities in the University of Zaragoza.



[redacted]

Figure 10.2: Evaluation of the speech-controlled automatized wheelchair

The results were satisfactory, but it was shown the difficulty that supposed the recognition of the disordered speech of the test subject. An added difficulty were the developmental disorders of the subject, who suffered of usual time-space disorientations when uttering the commands to the system as he could not relate the map shown on the screen of the computer as a representation of the evaluation scenario where he was placed.

Unfortunately, the early stage in which this thesis was at the time of the experiments did not enable the use of the proposed techniques for personalization, and only the baseline system was tested, which could not provide the best recognition results possible as the work during these years has shown.

10.1.2 Voice-triggered mouse and keyboard events

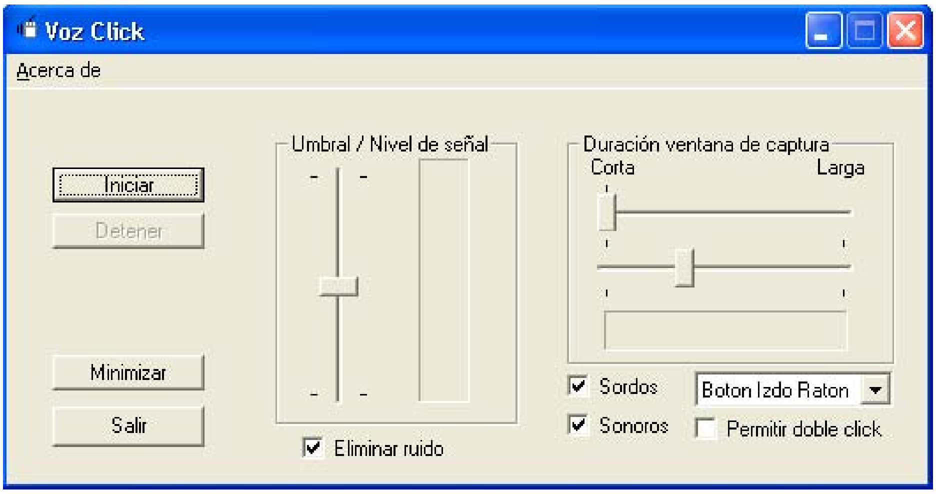
The control of different computer elements like the keyboard or the mouse can suppose a serious barrier for those people who suffer an affection over the fine motor system like people affected by cerebral palsy, as key or mouse strokes require a good control of the nervous system. The development of elements that can substitute these devices is of major interest for those who work for the handicapped community.

For many years, devices based on mechanical aids have been used for simulating keyboard or mouse events. Big-sized buttons that are handled by the user with different parts of the body like head, elbow or knee can be connected to the computer or directly to the mouse to trigger these computers events. However, these situations can be uncomfortable for the users due to the continuous repetitions of movements that the user has to do; or can not provide enough robustness as the rate of false activations raises with the involuntary or spastic movements of the user.

This situation led to the interest of the staff of Confederación Española de Federaciones y Asociaciones de Atención a las Personas con Parálisis y Afines (Spanish Confederation of Federations and Associations for the Assistance to People with Palsy and Related Handicaps)

(ASPACE)-Huesca for the development of an emulator of mouse and/or keyboard events controlled by voice or speech. The work was made within the collaboration framework between Coordinadora de Asociaciones de Personas con Discapacidad (Coordinating Committee of Associations of Handicapped People) (CADIS)-Huesca and the University of Zaragoza, funded by different institutions like the Diputación Provicial de Huesca (Provincial Council of Huesca) (DPH), the Instituto Aragonés de Servicios Sociales (Aragonese Institute for Social Services) (IASS) and the Caja de Ahorros de la Inmaculada (CAI). Several coordination meetings were made to define the settings and requirements of the application; which was given the name of "MouseClick" or "VozClick" [Saz et al., 2009a, Saz et al., 2009b].

The application processes the input audio signal captured by a microphone and applies an energy-based Voice Activity Detection (VAD) over it to determine when to trigger the keyboard or mouse event. The application was development for the Windows Operating System (OS) and presents an interface as shown in Figure 10.3. From this interface, the user or the person assisting the user can configure all the elements within the application: the energy threshold, the duration threshold and the sonority threshold.



☑ Eliminar ruido

☑ Sordos

☑ Sonoros

Boton Izdo Raton

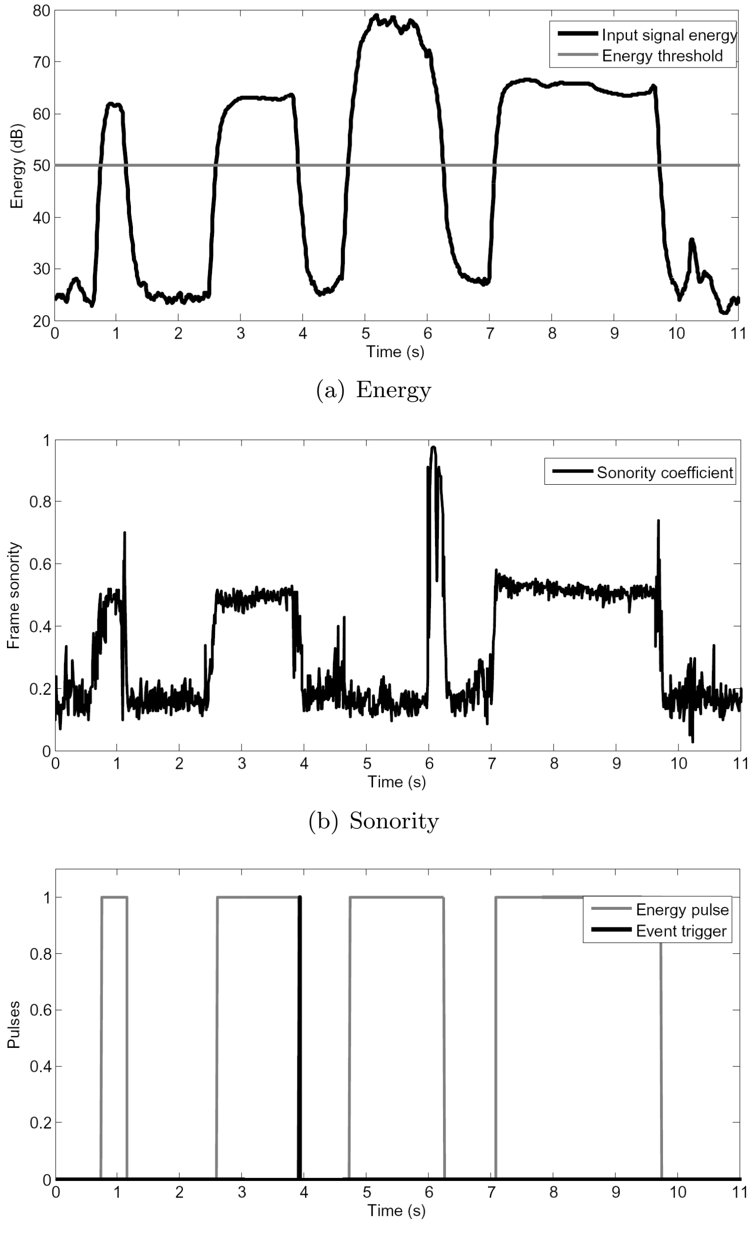
□ Permitir doble click

Figure 10.3: "VozClick" main window

The application monitors the input speech signal and looks for a speech pulse whose energy overpasses the defined energy threshold. When the pulse finishes it evaluates if the duration of the pulse has been in between the limits decided by the user (to avoid the activation with extremely short or extremely long emissions) and triggers the desired computer event (mouse stroke or key stroke). An extra threshold can be set to filter emission according to their sonority; this way, a user can decide to emit only voiced speech utterances (vowels or similar) or voiceless emisions (breathing or fricatives).

Example 10.1.1 An example on how "VozClick" reacts to different input entrances is shown in Figure 10.4. In Figure 10.4(a), the input energy for each speech frame is plotted, altogether with the energy threshold; Figure 10.4(b) shows the frame sonority value, which measures how sonorant the frame is according to the same speech processing methods followed in Chapter 5; and finally, in Figure 10.4(c), it is plotted the binary output of the energy thresholding and the trigger of the mouse event, as a binary signal too.

Four voice events were captured in the example signal. The first one was shorter that the minimum duration set to trigger the mouse event, while fourth one was longer than the maximum duration set to trigger the mouse event; hence, no one of them produced any output. The third voice event had the desired duration, but it did not match the sonority feature which was set to look only for sonorant parts of speech; hence, it was discarded too. Only the second voice event had the desired duration and sonority value to trigger the mouse event (black line in the bottom plot).



(a) Energy

(b) Sonority

(c) Output

Figure 10.4: Energy, duration and sonority thresholds on "VozClick"

The application was intended for the use with other software elements like slides, whose movement could be handled by speech or for complementing virtual emulators like a virtual keyboard or virtual mouse. The final evaluation was made with a Virtual Mouse ("Ratón virtual") device developed by the group led by Prof. Joaquín Ezpeleta of the Department of Computer Science and Systems Engineering of the University of Zaragoza in collaboration with the CPEE "Alborada" [Bergua, 2005, Bergua et al., 2006]. Figure 10.5 shows the evaluation experiment carried out during the media presentation of “VozClick” in Huesca (Spain). The test subject (a woman with spastic cerebral palsy) was able to emulate the mouse handling via speech, opening Windows applications or navigating the Internet.

This first subject and two more test individuals from ASPACE-Huesca with spastic cerebral palsy were testing the application during several months. First subject showed up a great ability with the voice interface and increased her efficiency with lower effort in the use of a virtual keyboard device (Click’n’Type) for the writing of documents in a Microsoft Word environment.

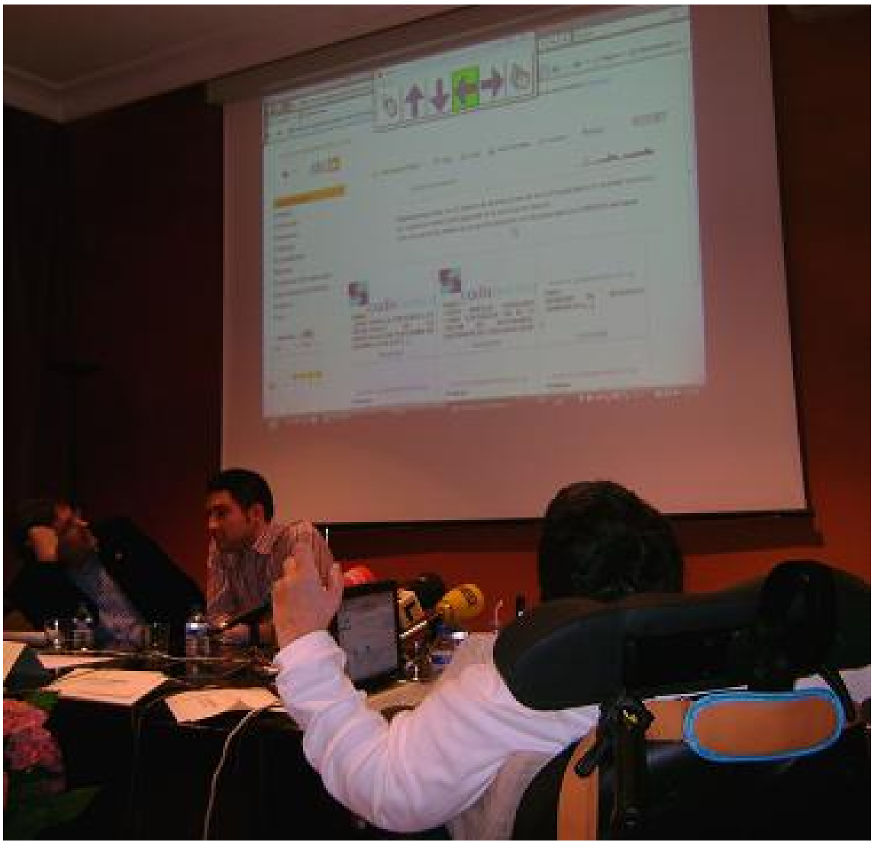


Figure 10.5: Evaluation of “VozClick” with the “Ratón virtual”

Further tests meant also the inclusion of new features in the application: Some of them affecting the interface of the application like a “Minimization” option that sends the application to the Windows tray to avoid the application window appearing constantly on screen and disturbing other elements; and some others affecting the audio processing within the application like the use of an spectral subtraction algorithm like Minimum Mean-Square Error Log-Spectral Amplitude (MMSE-LSA) [Ephrain and Malah, 1985] to reduce the input noise captured by the application due to different sources like electronic devices or environment noise. This presence of noise, especially electronic noise from the own computer, limits the possibilities of the application as the user has to raise the volume of speech to overcome this background noise level.

10.1.3 Speech control of virtual mouse

The final implementation oriented to applications for the handicapped was the inclusion of the state-of-the-art ASR system used in this thesis as an interface in the virtual mouse device (“Ratón Virtual”, mentioned previously) generated by Prof. Joaquín Ezpeleta of the Department of

Computer Science and Systems Engineering of the University of Zaragoza in collaboration with the CPEE “Alborada” [Bergua, 2005, Bergua et al., 2006].

The latest work in this line of research supposed the creation of the “MICE” environment for introducing new drivers in the control of the mouse device [González, 2009]. One of the possibilities that was considered was the inclusion of an speech interface based on the baseline ASR system presented in this thesis. The environment provided the tools in Java Native Interface (JNI) to communicate the ASR libraries in C with the virtual mouse device in Java.

This work is currently ongoing and is requiring an initial evaluation of the performance of the system and the tuning of the baseline system towards the final inclusion of the tools that can provide an adaptation of the system to the user, as this thesis has showed that this is strongly necessary when developing speech systems for the handicapped with possible speech impairments.

The demand for this kind of devices is high, as it can really provide enhanced communication to severely handicapped individuals, so this opening line of work would strongly count with the help of institutions like the CPEE “Alborada” or ASPACE-Huesca.

10.2 CASLT Tools - “Comunica”

“Comunica” appeared as the joint effort from the VIVOLAB speech group at the I3A and the therapists, educators and staff of the CPEE “Alborada”, trying to bring together the experimental research in speech technologies with the development of real-world CASLT tools [Rodríguez et al., 2008a, Saz et al., 2008b, Saz et al., 2009h, Rodríguez et al., 2009]. It aimed to cover the full process of language acquisition, from the phonatory skills to the requirements of the functional language, through phonological acquisition.

10.2.1 “PreLingua”

“PreLingua” is an application for the stimulation and training of basic phonatory skills in infants or handicapped persons with severe voice pathologies. The application was developed by William Ricardo Rodríguez within “Comunica” [Rodríguez et al., 2007, Rodríguez et al., 2008b]. Its basis are speech processing to extract acoustic features like intensity, fundamental frequency or formants.

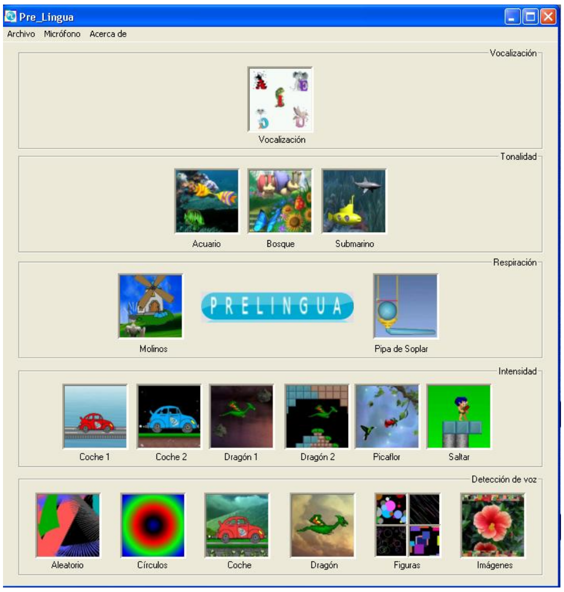
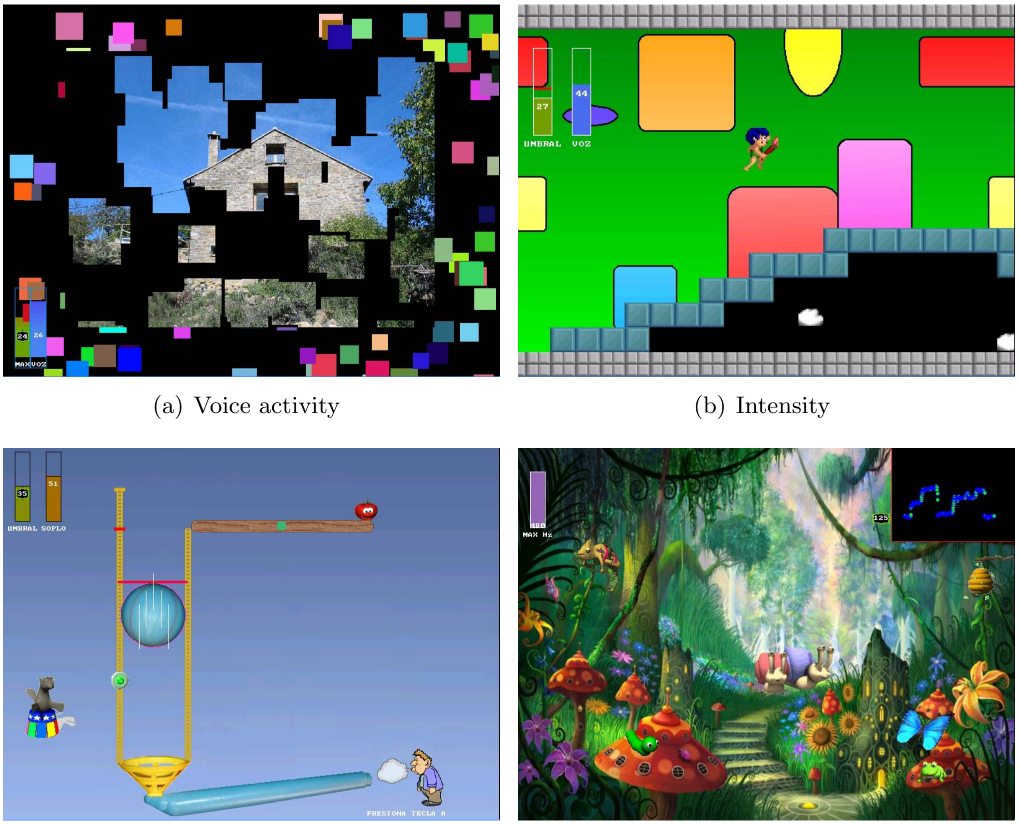


Figure 10.6: “PreLingua” main window

The interface of the application is based on the main window shown in Figure 10.6; in this window all the activities of the game are gathered in categories and can be accessed directly with just one mouse click.

The activities were organized in a pyramidal structure from the more basic categories to the more complex elements of voice:



(a) Voice activity

(b) Intensity

(c) Respiration

(d) Pitch

Figure 10.7: Activities in "PreLingua"

The bottom of the pyramid is occupied by the *voice-activity detection activities*. These activities were intended for the stimulation of the speech production in infants with different disabilities (i.e. hearing disabilities). As the patient utters any sound, the screen shows an image moving or changing shape like it can be seen in Figure 10.7(a).

Next step in the pyramid are the *intensity activities*. Further than just distinguish between presence or absence of sound, these activities stimulate the sense of volume or sound intensity associated to the patient's oral production. Figure 10.7(b) shows one of these intensity activities, where a character advances in a circuit or labyrinth as the patient produces and oral input above a certain intensity level.

The further step in the application are *respiration activities*. A correct control of breathing and airflow is requested for the correct production of speech with a correct and sustained volume. The respiration activities like in Figure 10.7(c) simulate a blowpipe or windmills that are activated by the intensity of the patient's voice, but only in the presence of unvoiced segments of speech; this is, segments where fundamental frequency or pitch is not detected.

Final step of voicing stimulation in the pyramid are the *pitch activities* like Figure 10.7(d) to help the patient to modulate and control the fundamental frequency of the speech production. A steady pitch production is a measurement of good voice quality and the activities aim to provide feedback to the patient by means of a character that navigates up and down the screen according to the pitch value of the patient’s voice.

In the top of the pyramid, the *vocalization activity* helps the student utter the first full vocalic sounds with the help of a representation of the Spanish formant map where the user has to match the vowels production to the standard values plotted in the application. An important work was done in this area to make this activity robust to children’ speech, as high-pitched individuals produce errors in the formant detection. Moreover, children present formant values different from the standard values of formants in adults as their vocal tracts are still growing, so the normalization of the formant values is required in this case [Rodríguez and Lleida, 2009].

10.2.2 “Vocaliza”

“Vocaliza” was the first CASLT tool developed within “Comunica”, in a work carried out mainly by Carlos Vaquero [Vaquero, 2006]. “Vocaliza” [Vaquero et al., 2006, Vaquero et al., 2007, Vaquero et al., 2008b, Vaquero et al., 2008a] is a CAPT tool that also trains the semantic and syntax levels of language as well as the phonological level.

The main window in “Vocaliza” is seen in Figure 10.8 and allows to access directly to all functionalities of the game, including: Creation and management of user profiles, introduction of new activities, enrollment phase and the 4 designed activities.

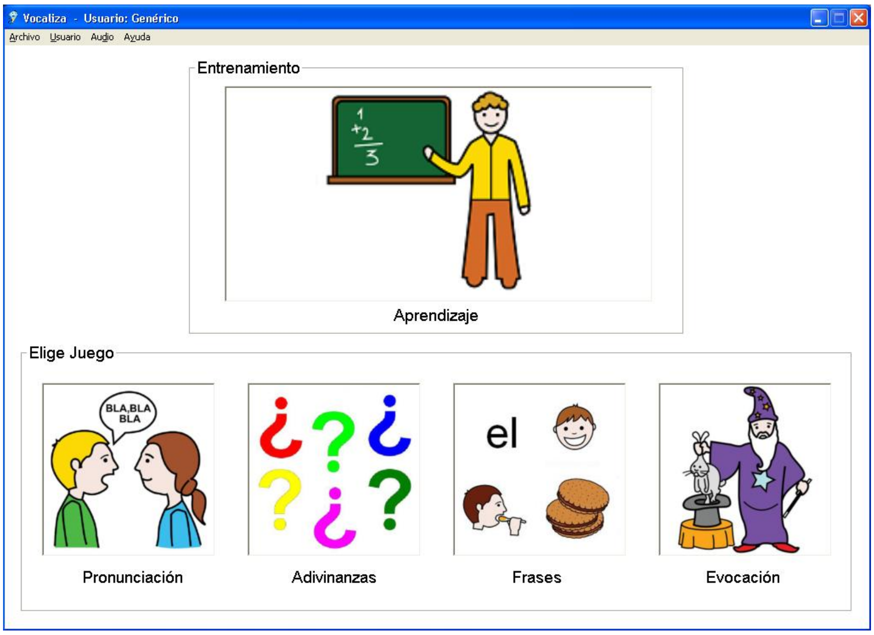


Figure 10.8: “Vocaliza” main window

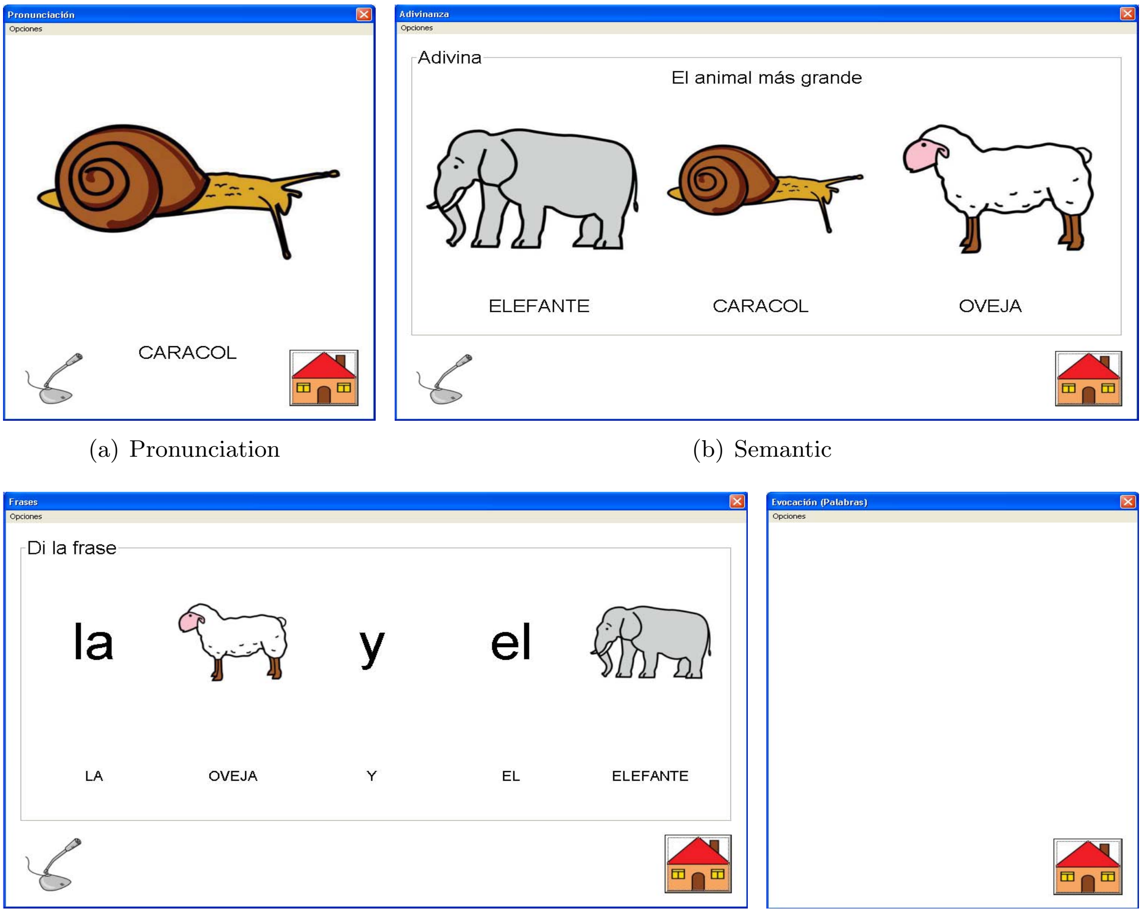
The four activities designed for “Vocaliza” focused on the training of the speech of the student by providing an audio-visual feedback of the speech quality of the student’s speech.

Pronunciation activities like in Figure 10.9(a) present a word to the user, awaiting for the utterance of that word by the student. The system runs ASR over the student’s input utterance to decode whether the speaker is pronouncing the prompted word and posteriorly runs a word-based UV system to give a mark to the student’s pronunciation.

Semantic activities like the one seen in Figure 10.9(b) give the student a riddle game, with a question and three possible answers, and the student has to orally answer the riddle. The system runs ASR over the student’s utterance aiming to recognize the correct answer to the riddle. When the correct answer is given, a mark is provided to the user relating to the number of trials that the student has done to pronounce the correct word.

Syntax activities like in Figure 10.9(c) present a sentence to the user, awaiting for the utterance of that sentence by the student. After ASR recognizes the prompted sentence from the user, the UV system provides a mark about the overall quality of the student's pronunciation.

Evocation activities like the one presented in Figure 10.9(d) await for a word uttered from the speaker to present the word on screen. When the ASR system decodes one of the words in the vocabulary, it presents it in the screen. Evocation can help to stimulate children's oral language by the cause-effect stimulus created with the apparition of the mentioned image.



(a) Pronunciation

(b) Semantic

(c) Syntax

(d) Evocation

Figure 10.9: Activities in "Vocaliza"

An important point in the application is the use of AAC technologies. Tentative users of the application might be children with different impairments, including sensory impairments that would limit their possibilities of interaction with the software application. For this reason, the presentation of the activities and the feedback provided after each activity are based on three different elements: Image, text and audio.

Images are the basis of the AAC systems. In "Vocaliza", they are the center of the all interaction of the student with the tool. In Figure 10.9, it can be seen as images appear in all the presentations of the activities, providing all children, who might suffer different development disorders, an easy way to understand what is asked from them. Audio reinforces the correct pronunciation of the presented word or sentence; with the use of audio, an extra stimuli can be provided to different users, while it can be removed for the hearing impaired ones. One step further, the audio can be synthesized via the Lernout & Hauspie TTS3000 or can be recorded directly from the therapist's speech. The selection of the audio reinforcement (synthesized or pre-recorded) is very important and can be decided by the therapist according to the needs of each user, as it has seen widely seen how different users require different types of audio prompting [Saz et al., 2010a]. Finally, text can provide an extra stimuli to many users, while it can also be removed for children with vision impairments or lecture difficulties.

The technologies embedded in "Vocaliza" are ASR to decode the utterance from the speaker, UV to provide a confidence measure once the correct answer has been decoded and MAP speaker adaptation to create acoustic models adapted to the speaker speech.

The ability of "Vocaliza" to provide a coherent feedback to the user is based on the properties of the ASR within the application to correctly recognize or reject the utterance according to the quality of the pronunciation. As it was shown in Section 4.6, the recognition rate of the ASR system decreases heavily as the speaker produced more mispronunciations in the utterance. In these cases, "Vocaliza" keeps forcing the speaker to utter again the word or sentence until it is more accurately pronounced by the student and is recognized in the ASR. The quality measuring system is based on a UV algorithm based in [Lleida and Rose, 2000] where two different models, one trained from unimpaired speech and one from impaired speech are used to calculate a measure of this quality. Finally, speaker adaptation via MAP helps create SD acoustic models as it has been shown that they help to increase the recognition accuracy in ASR systems and the detection accuracy in PV and UV algorithms.

10.2.3 "Cuéntame"

"Cuéntame" was developed within "Comunica" by Antonio Escartín [Escartín, 2008] and aims to stimulate the functionalities of language in individuals with language impairments. The language levels trained by "Cuéntame" are the next step in language production and relate to the capabilities of the student to use language to describe things, or interact with the environment.

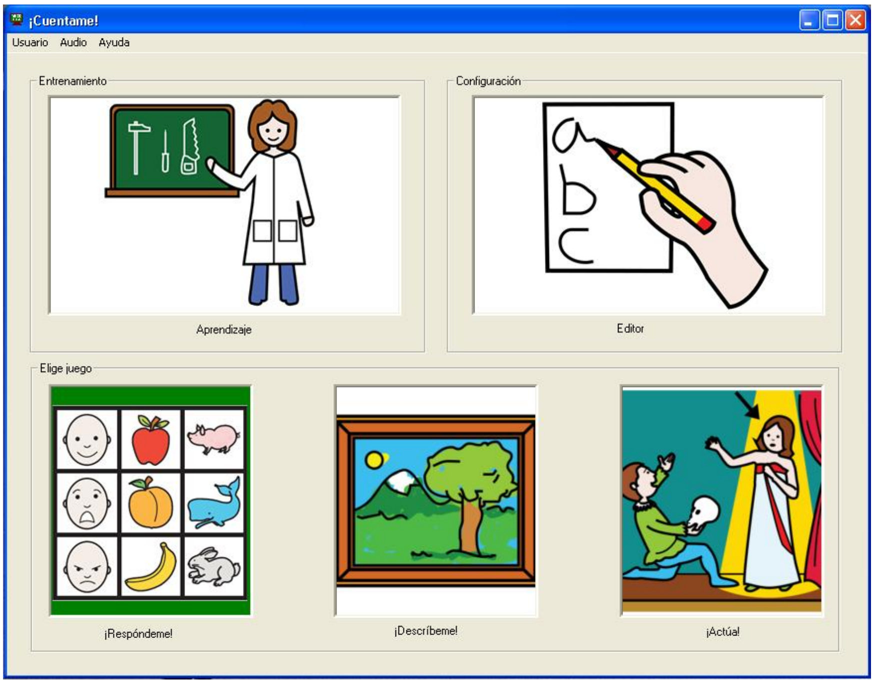
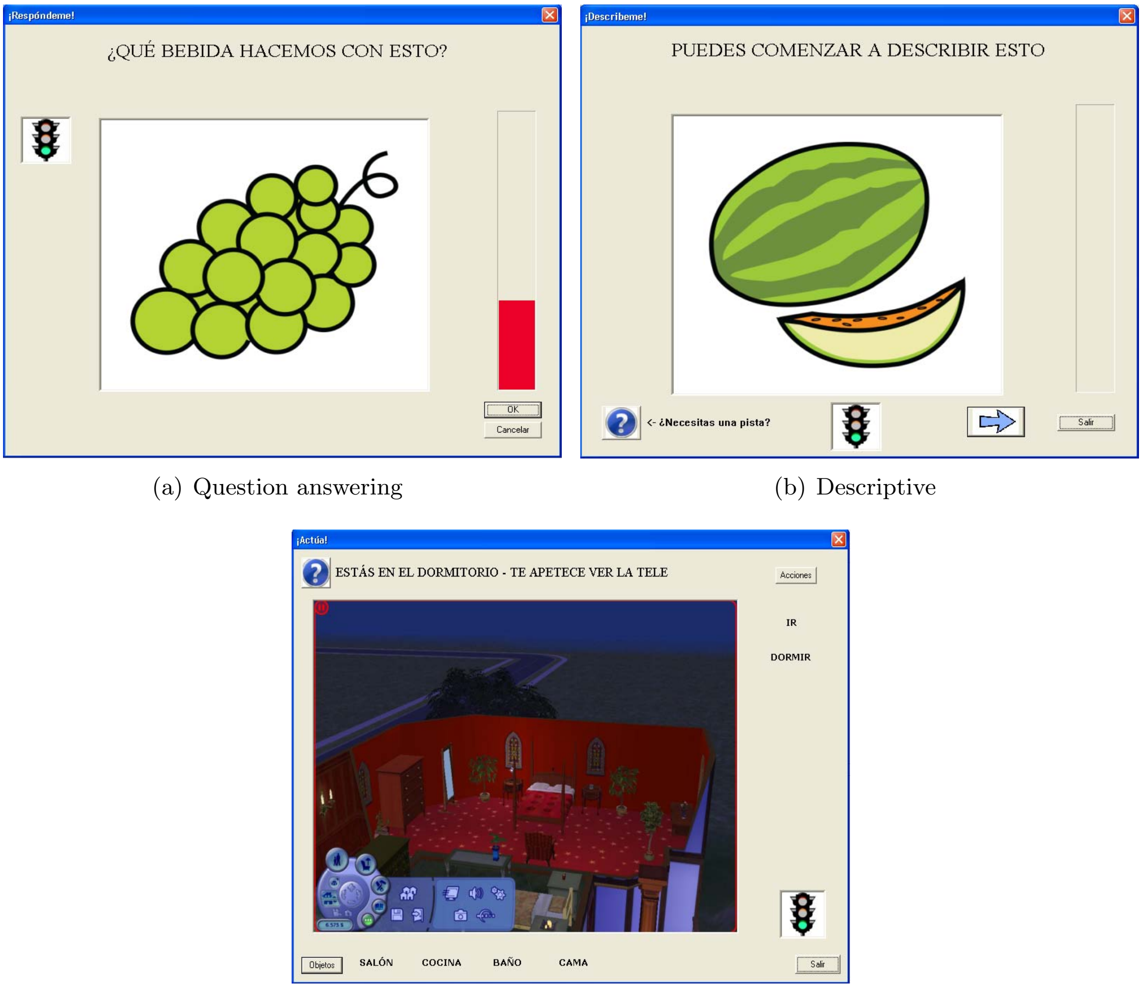


Figure 10.10: "Cuéntame" main window

As in all the applications in “Comunica”, the main window of “Cuéntame” in Figure 10.10 allows the access with a single mouse click to all the functionalities of the application: user configuration and activities.

Three activities were designed for “Cuéntame”, aiming to cover three aspects of the functional language.



¿QUÉ BEBIDA HACEMOS CON ESTO?

(a) Question answering

PUEDES COMENZAR A DESCRIBIR ESTO

<- ¿Necesitas una pista?

(b) Descriptive

ESTÁS EN EL DORMITORIO - TE APETECE VER LA TELE

(c) Interactive

Figure 10.11: Activities in “Cuéntame”

Question answering activities like in Figure 10.11(a) present an open-answer question to the student who has to answer with a whole sentence. An ASR system is run to decode the input utterance, seeking for a set of keywords that define how accurately the student has approached to the full correct answer.

Descriptive activities like the one presented in Figure 10.11(b) stimulate the student to utter whole sentences describing the object appearing on screen according to some aspects (color, shape, etc) predefined by the therapist. The ASR system again looks for the keywords that mark the approximation to a predefined full correct sentence that describes the object.

Interactive activities like in Figure 10.11(c) present the student with an scenario and a goal to accomplish and the student has to navigate orally through different stages and actions to finish the proposed goal. The student has to match pairs of ‘action-object’ that are recognized by the ASR system and reacts to the proposed action of the student.

The technologies embedded in "Cuéntame" are ASR to decode the oral utterance from the speaker, semantic-syntax analysis to infer the possible correct sentences that the speaker can utter and UV to discard OOV words. The main feature that distinct "Cuéntame" from a purely CAPT tool like "Vocaliza" is the fact that "Cuéntame" motivates the student to utter full meaningful sentences, instead of just isolated words or prompted sentences. The difficulty of the language modeling, as the therapist can include as many activities as desired who might vary largely from each other, made that the ASR runs a keyword spotting on several words that mark the correctness of the uttered sentences (noun, verb, pronouns, adjectives, etc) to decide whether the student has approximate to the correct full answer. Furthermore, treatment of OOV words is made via this keyword spotting that rejects all words not included in the vocabulary of the activity and an extra UV to discard any of these words that may have been recognized as an in-vocabulary word.

10.2.4 Other tools: "ReFoCas"

One of the main features of "Comunica" has been the continuous communication with users and therapists who used the applications to know their opinions about the tools, their problems (if existed any) with them and their ideas for further improvements or developments.

From this feedback, the idea of providing an automated tool for the phonological register of students in different education institutions appeared by proposal of a group of speech therapists (Belén Gómez, Rebeca Sampedro, Sandra del Río, Ana María Escáriz and Ana Isabel Costa) from different schools from A Coruña and Pontevedra (Spain).

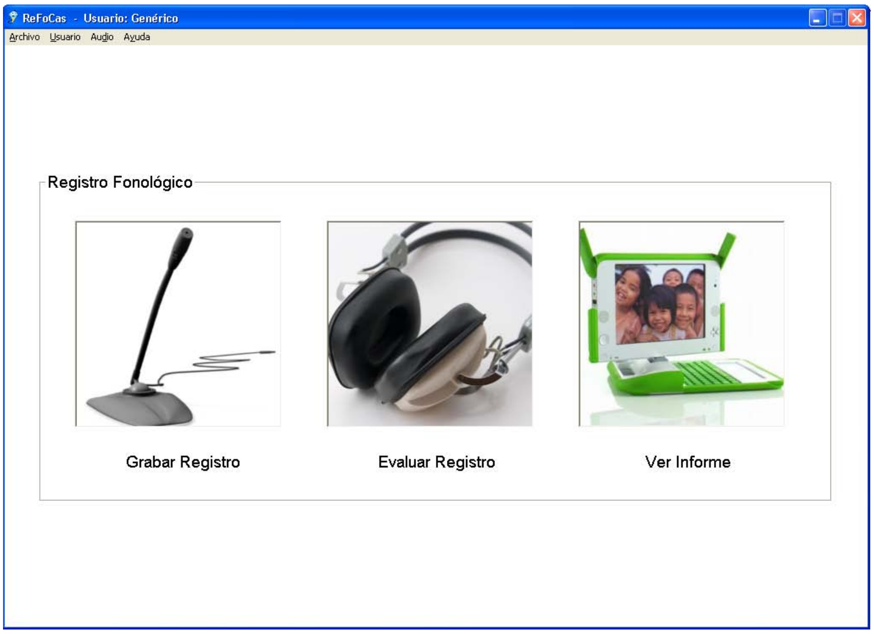


Figure 10.12: "ReFoCas" main window

This was the origin of "ReFoCas" and its Galician translation "ReFoGal". The tools included a set of words defined by the therapist to assess the ability of the students in different phonological contexts (singles phonemes, consonant clusters and vocalic groups). The development of the tool meant using the enrollment phase embedded in "Vocaliza" to capture the utterances from the student and creating an interface that allowed the therapist to review each utterance as many times as needed prior to evaluate the quality of the pronunciation. With the complete evaluation of all the utterances, a report is created in a printable form for a posterior use by the therapist.

The main interface of “ReFoCas” is presented in Figure 10.12, where the three stages of the phonological register can be seen: “Grabar Registro” (Record Register) presents the prompts of the different words chosen and performs the recording process of the selected words; “Evaluar Registro” (Evaluate Register) allows the therapist to review the recorded utterances and evaluate the different phonemes or parts of speech desired; finally, “Ver Informe” (See Report) shows the automatically generated report in a printable format.

This tool did not make use of any speech technology like ASR or PV, as at this point the tool only intended to be an interface for the process of manually labeling the student’s speech. However, it aimed to fulfill a need of many speech therapists and it helped to open the range of collaborations with education institutions in “Comunica”. Furthermore, this initial translation of the interface in “Vocaliza” to a language different that Spanish, Galician, can be the beginning of the development of novel tools in other languages. In the future, novel possibilities for semi-automated phonological registers could be developed, where the tool could make an initial evaluation of the word or phonemes to be reassured or corrected by the therapist.

10.3 L2 Tools - “VocalizaL2”

The three tools defined initially in “Comunica” were designed, as mentioned, to cover all aspects of language acquisition in children with possible learning difficulties or language delays. As further possibilities for the development of new tools arose, some of them were carried out in joint work with different persons and institutions.

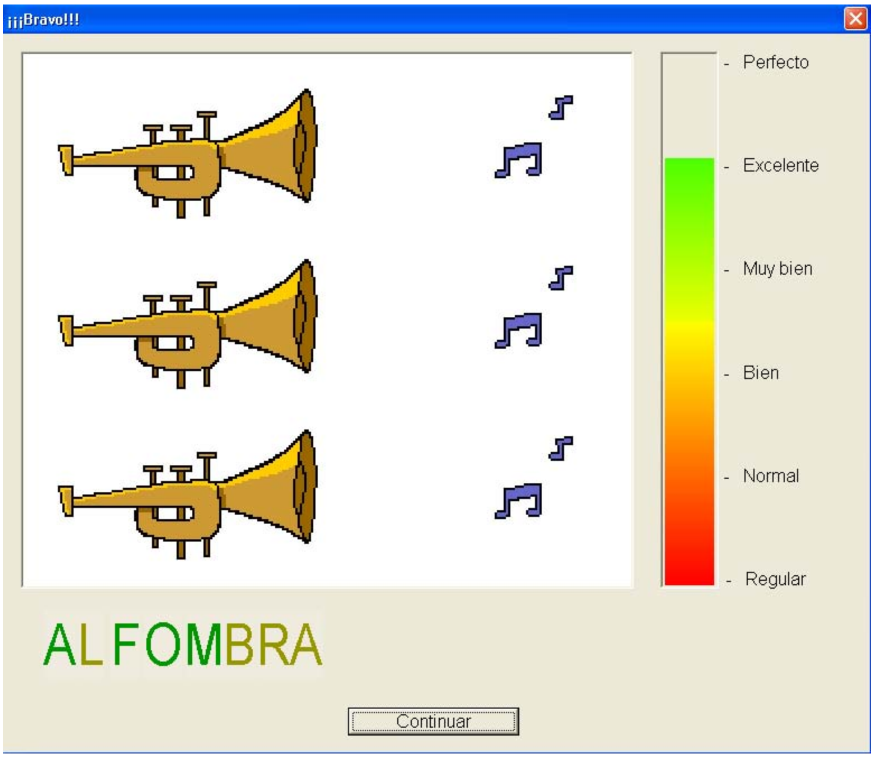


Figure 10.13: Feedback provided in “VocalizaL2”

“VocalizaL2” was developed as an expansion to “Vocaliza”. It used the same interface options than “Vocaliza” in terms of user interaction, AAC technologies and user profiles, but providing a different feedback than the original application. While “Vocaliza” provided a simple word-level or sentence-level verification outcome of the speech abilities of the user in the different activities, "VocalizaL2" used the normalization method for detection of phonetic mispronunciations presented in Chapter 8 to provide a feedback on the phonetic skills of the user. This feedback is provided as seen in Figure 10.13 where each phoneme of the word is shown within a 3-colour scale (green: correct, yellow: average, red: poor) according to the final score achieved by the phoneme after the normalization in the PV algorithm. A word-level feedback is provided as in "Vocaliza" as an average of the marks given to all the phonemes in the word.

Due to its extended ability to provide phonetic feedback, the application was initially oriented to L2 learning [Rodríguez, 2008, Saz et al., 2010b], as L2 users require an improved phonetic feedback as this new tool could provide. To evaluate the pedagogical and technical possibilities of the application, it was tested in a real-world experience within a multilingual environment at the Spanish classes taught by Victoria Rodríguez in the Vienna International School (VIS) in Vienna (Austria).

10.3.1 Experience with "VocalizaL2"

The experience was carried out during a period of 5 non-consecutive weeks, where a 45-minute session with the application was run weekly. During each session, every student could practice with the application for a time of approximately 10 minutes in one of the 2 computers in which the application was installed. The experimental group of students was composed of 12 students in their 6th grade of education (11 years old). In this grade, students were starting the learning of their third language, so all of them were beginners in Spanish. In the group there were 8 boys and 4 girls and the distribution in terms of the mother tongue of every student was as seen in Figure 10.14. Mother tongue and the interlingua that students were using to interact with the new language (classes were taught on English) may have an important role in the understanding of how different students reacted to different pronunciation learning issues.

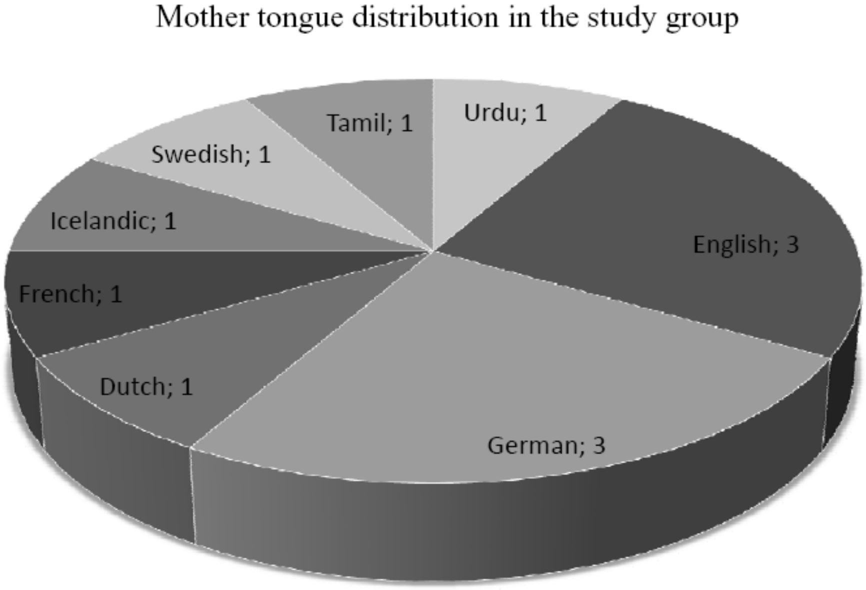


Figure 10.14: Mother tongue distribution in the experimental group

The set of words was chosen according to the vocabulary studied by the children in the classes and divided among the 5 sessions. Ten words were finally chosen for each session, and each student pronounced 2 times each word. Session 5 had a different approach and was designed as a gameplay in which students were divided in teams and they pronounced alternatively the words; only 7 words were, hence, programmed for this session. The words in the sessions were:

- Session 1 (Food): beber, bocadillo, botella, carne, cereales, cerezas, chocolate, galletas, hamburguesa, helado.
- Session 2 (Daily routine): andar, bañarse, cocinar, dormir, ducharse, escribir, jugar, llorar, pegar, trabajar.
- Session 3 (Animals): araña, ardilla, burro, cerdo, conejo, foca, gallina, mariposa, pájaro, rinoceronte.
- Session 4 (At home): armario, cocina, cuadro, bañera, escalera, frigorífico, sillón, ventana, espejo, librería.
- Session 5 (Neighborhood): acera, ambulancia, balcón, calle, calzada, farola, papelera.

The application offered the possibility of giving pictorial, auditive and written prompts to the user. The three options were chosen during sessions 1,2,3 and 5. In session 4, the written prompt was omitted to observe how students dealt with only auditive prompting.

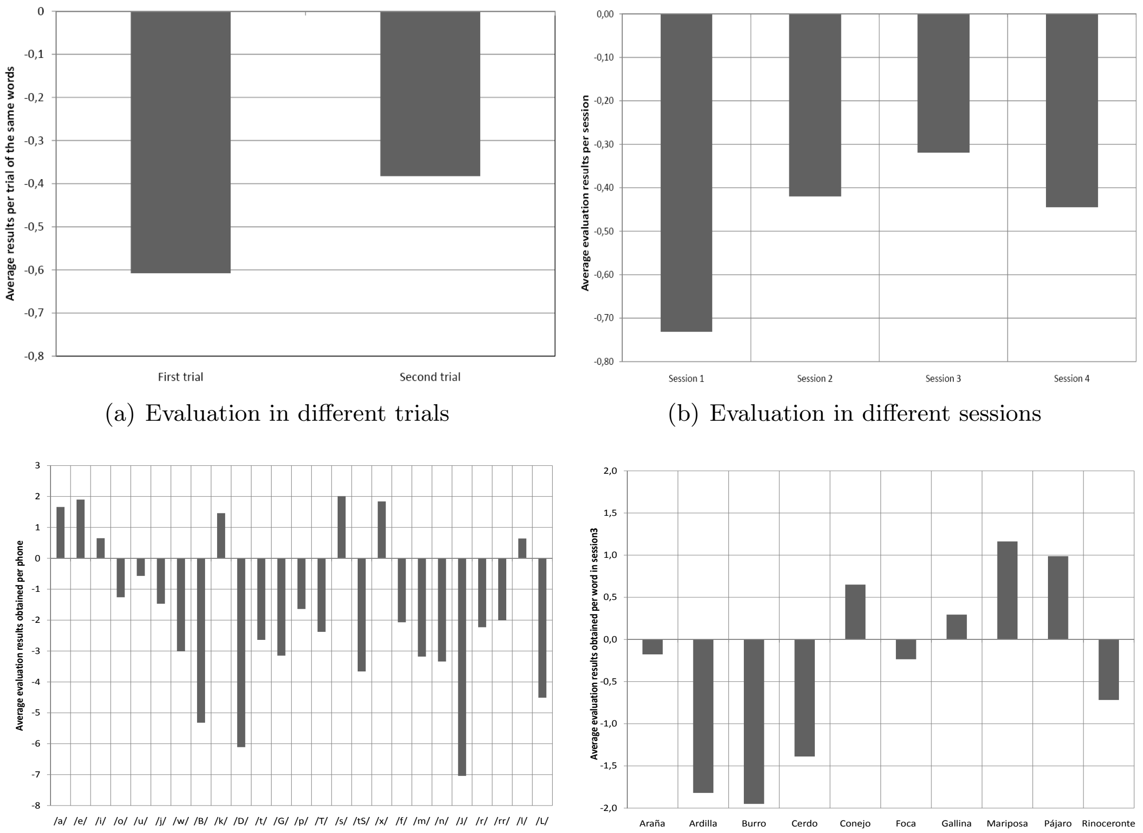
The experience proposed in the experiment had two different aspects to evaluate. On one hand, the pedagogical results of using a computer-based tool for L2 learning in the proposed environment; and, on the other hand, the evaluation of the tool and its ability to correctly verify the pronunciation of the users.

The opinions of the students with the application were collected afterwards each session by their teacher. Students were aware, despite their short age, that the application was providing an evaluation coherent with the effort and interest they put on their pronunciation; so they strongly tried to improve in different trials and sessions. They evaluated positively the interface and felt really motivated to have more classes with the application, although they also pointed out some weak points like the lack of naturalness of the synthetic voice that provided the audio prompt and the sometimes odd evaluation results given by the tool (possibly due to the presence of noises or disfluences).

Considering the different possibilities provided by the application to L2 teachers, the teacher realized that session 4 without the written prompt was more challenging for students, as they had to rely only on the audio prompt, which could be more interesting for advanced students. Moreover, session 5 was seen positively by all students and teacher, as gameplay activities provided an extra motivation to young learners.

The initial approach of the experience was to make an evaluation of the performance of the tool reviewing the evaluation results given by the tool to all speakers. Unfortunately, no labeled data was available from the speakers and the short number of sessions made difficult a more precise study of the results. Hence, the performance evaluation was made studying the log-likelihood scores that the application assigned to each utterance and that were kept stored within the application. Three points of relevant interest were studied: Evaluation in different trials of the same word, trend of the evaluation through different sessions, and specific results for different phonemes and words [Saz et al., 2009f].

Two different trials of every word were programmed in the sessions. Average results obtained by the students in the first and second trials of every word are provided on Figure 10.15(a). A review on the results separating both trials showed that the evaluation marks given on the second trial were higher than in the first trial. This was consistent with the fact reported by the teacher that students put a bigger effort in the second trial after they got the evaluation of the first trial; furthermore re-prompting for the second trial reinforced the correct pronunciation of the word in the students and helped them to improve their pronunciation.



(a) Evaluation in different trials

(b) Evaluation in different sessions

(c) Evaluation for different phonemes

(d) Evaluation for different words

Figure 10.15: Evaluation results in the “VocalizaL2” experience

Also, the average results obtained by the students in each session are shown in Figure 10.15(b). These results showed an important improvement in the evaluation obtained from session 1 to session 2, followed by a lesser improvement in session 3 and a reduction in the results in session 4. Improvement from session 1 to session 2 could be explained by the fact that the children got used to they way in which application worked and they understood it better, putting more effort in their pronunciation. Results in session 4 (slightly worse than previous session), might corroborate the fact pointed out by children and teacher that uttering words without the written prompt was harder for students.

Finally, a study of the performance of the ability of all speakers to pronounce different words and phonemes was performed. The average log-likelihood results for all the 25 units used (23 phonemes plus allophones [*j*] and [*w*]) for all speakers and sessions are shown on Figure 10.15(c). Lack of labeled data diminished the ability to extract conclusions, but trends were similar to the trends of mispronunciations in young Spanish children in the natural process of language acquisition [Bosch-Galcerán, 2004]. Students achieved higher marks on vowels, while special phonemes like /*tS*/, /*L*/ or /*J*/ received lower marks. Also noticeable was the significant worst results achieved in the glides /*j*/ and /*w*/ compared to the corresponding vowels /*i*/ and /*u*/; even if the sound is similar, it is usually noticed how glides in diphthongs are more difficult to pronounce than the vowels for young Spanish children [Bosch-Galcerán, 2004].

A study over the results given to a set of words was also made, the 10 words studied were the words in session 3 as it was the session in which the students were more used to the application. The average results for all speakers with these words are shown on Figure 10.15(d). Again, lack of labeled data made unable a comparative study, but some trends could be observed that again agreed with theories of phonetic acquisition in Spanish. Words *burro* (SAMPA: */burro/*), *ardilla* (SAMPA: */ardiLa/*), *cerdo* (SAMPA: */Terdo/*) and *rinoceronte* (SAMPA: */rrinoTeronte/*) were the worst pronounced by the students, which was consistent with the fact that these words contained the vibrant phonemes /rr/ and /r/, which is usually a difficult feature for learners of Spanish, especially in the case of /r/ in coda position.

Further work arose as result of this experimental work. In terms of user interface, a major effort should be done to adapt the application environment to an interface more suitable for possible adults learners, as this population is very willing too for the use of L2 tools. A more natural synthetic voice would be required to make the tool more useful, as the audio reinforcement was seen as very relevant and attractive for tentative users of L2 technology, who may see their learning possibilities reduced with a poor TTS voice.

Chapter 11

Discussion

Tiger got to hunt
Bird got to fly
Man got to sit and wonder: "Why, why, why?"
Tiger got to sleep
Bird got to land
Man got to tell him he understand

The Books of Bokonon c. 81

-Kurt Vonegut, *Cat's Cradle*

11.1 Introduction

Although several elements have already been discussed in their respective Chapters of this thesis, this Chapter wants to raise up the discussion in some elements which require an eagle-eye view over different Chapters altogether or that may be of special relevance for the objectives and procedure of the thesis. With all this, there will be three discussion points which will relate to the three main aspects that this work has tried to cover:

- The understanding of the origin of the speakers' disorders will come from the knowledge acquired in the analyses in Chapters 5 and 6; with Section 11.2 analyzing all the hypotheses and how the acoustic and lexical aspects of these disorders have influenced a baseline speech-based system like the ASR in Chapter 4.

- The different issues that affect the designing of personalization strategies to avoid the pernicious influence of these disorders will be discussed in Section 11.3. Special attention will be paid to how two different tasks like ASR and PV shown in Chapters 7 and 8 react to the personalization techniques in the presence of severe mispronunciations and acoustic degradation in the speakers, with a final remark on its influence in the proposal for unsupervised systems in Chapter 9.

- Finally, Section 11.4 will point out whether it is possible or not to translate directly all this theoretic knowledge to the development of aids for the handicapped in Chapter 10. This discussion wants to focus on how different actors (developers, users, public agencies) have to work and collaborate together to fulfill these objectives.

11.2 Origins and Relevance of the Studied Speech Disorders

One of the points where more effort has been put during this thesis has been the acquisition and analysis of a full corpus containing disordered speech from a total of 14 young speakers. This corpus has been the basis for all the experiments carried out during this thesis and it is hoped that more research can be conducted over this corpus or future expansions of it.

Since the very beginning, it was understood that it was not useful to perform any kind of experimental framework without trying to understand, at least minimally, the affections of the speakers' impairments in their speech and their origin. For that reason, a lot of literature review in the early stages of this thesis' work was carried out in the areas of speech impairments, phonetics and phonological acquisition by young learners. This provided some clues on what was to be expected from this speech and justified the perceptual lexical labeling carried out by the set of human experts, as in the end it was seen that it was at the lexical level where the speakers confronted more difficulties.

The dramatic effect of this disordered speech on the shown ASR results, for instance in the TI and TD cases in Chapter 4, has shown how these disorders produced a major loss of performance in these systems.

This effect of the speakers' disorders over their speech could be separated in two causes: Acoustic degradation and lexical degradation:

- Acoustic degradation appears in those cases where the speaker presented a certain articulatory difficulty which blurred, distorted, or degraded the correct pronunciation of a phoneme or sound. It happens, as explained in the Introduction, as the consequence of morphological impairments or neurological impairments that limit the motor abilities of the patient, including the phonatory/articulatory organs. The measurement of this acoustic loss of quality was made in Chapter 5.

- Lexical degradation appears in those cases where the speaker produced a different phoneme instead of the expected phoneme or sound. The origins of lexical degradation can be many: Articulatory difficulties lead to the change of some of the articulatory properties of the phoneme, or the speaker has a difficulty to distinguish phonologically the sounds at the cognitive level. The characterization of this lexical variants was performed in Chapter 6.

This discussion will consider a simplification, assuming that both effects are fully separable in the speakers; this is, the processes that lead to acoustic degradation (i.e. dysarthria) are totally different to the processes that lead to a perceptible lexical degradation (i.e. acquisition delays). Although it is understood that all these factors, and many others introducing more variability, are in the end intermingled; this approximation is necessary to reduce the complexity of the discussion.

With this separation, Figure 4.10 makes sense in how it tried to fully separate the influence of the acoustic and lexical variability caused by these disorders in the ASR results. The large difference between the results in the 52.29% of words who did not contain any lexical mistake (as indicated by the human experts) and the 47.71% of the words who did indeed contain lexical inaccuracies (9.1% vs 49.24%) marked the higher prevalence of the lexical disorders over the acoustic disorders in these speakers. This was coherent with the impairments of the speakers, where cognitive disorders (Down's syndrome in many cases) were suffered by most of the speakers and this is rather associated to lexical difficulties than to acoustic difficulties, as the functional modules of language are more affected than the morphological generators of speech, although some of the speakers suffered degrees of dysarthria or dysphemia.

These impressions and results were assured with the acoustic and lexical analysis of the 14 speakers' speech carried out in Chapters 5 and 6 respectively. The acoustic degradation in their speech, although noticeable and measurable, was not so high as to produce a total confusion in the analyzed features. On the contrary, the study of the lexical properties in the corpus brought many more elements for discussion, especially in their implications with the process of language acquisition in these speakers.

It is for this reason that lexical adaptation has managed to show some impressive results with that subset of speakers who obtained a medium-high WER results in the SI results. For these cases, speakers *Spk*02, *Spk*07, *Spk*08, *Spk*09, *Spk*10, *Spk*12, *Spk*13 and *Spk*14, the improvement caused by lexical adaptation was up to 30-40%. These were also the results seen during the study in Section 7.2.1 with Speakers *Spk*07 and *Spk*08 on the amount of adaptation data for lexical adaptation. These speakers, who are suffering a big deal of lexical variability in their speech are, hence, taking advantage of the new SD lexicons. The drawback is, unfortunately, that speakers with lower disorders and less lexical impact, are obtaining smaller improvements or even losses for incorporating the new lexical variants. Hence, a selection method of lexical variants could be necessary before incorporating them directly in the new SD dictionary.

One of the implications of this prevalence of the lexical disorders in the speech used for research in this thesis is that the results arising from it can only be considered for similar speakers with similar disorders. The combined approaches that make use of acoustic and lexical adaptation might not work with similar performance in speakers with a bigger influence of acoustic disorders in their speech (the case of patients with cerebral palsy and dysarthria). The same affects to the effectiveness of the PV algorithms developed in this thesis, which have shown a decent performance to detect lexical mispronunciations, but it was not researched how the acoustic variability induced in the speech signals by the speakers' disorders will affect this performance, or how this deal of algorithms would be reliable as a measure of speech quality in lexically correct speech.

11.3 Issues on Personalization of Speech Systems for the Handicapped

The bigger effort of the thesis has been, as seen, the study of several issues and schemes for providing an improved recognition performance through speaker adaptation and personalization. All these issues can be categorized for analysis in 3 different categories: acoustic vs lexical analysis in two different tasks, ASR and PV; and, approximations to the unsupervised on-line situation.

11.3.1 Acoustic vs lexical adaptation for ASR

One of the major interest points that this thesis has faced regarding adaptation techniques has been the joint use of acoustic and lexical adaptation. Table 7.1 showed how the use of a standard acoustic adaptation technique where the baseform transcriptions were fed to the MAP framework lead to an improvement of around 49%; while on Table 7.10, it was seen that the combination of an acoustic adaptation which only adapted to the lexically accurate data and lexical adaptation achieved an improvement of 45%. Although the approaches are different, the results are similar; but this can be argued to occur because they are two nearly-optimal approaches.

Example 11.3.1 An example of how different types of adaptation can work in the presence of disordered speech can be illustrated by Figure 11.1. This example shows two utterances of the word 'árbol' by Speaker *Spk*05 from the disordered speech corpus. The experts accorded that in both utterances the sounds [r] and [l] had been deleted from the speaker's pronunciation, reducing the baseform transcription [arbol] to [abo].

In the purely acoustic adaptation, the MAP algorithm uses a Viterbi forced alignment to decode the phoneme boundaries as shown in the upper left signal, and each segment is used for adapting the corresponding unit. In this case, phonemes [r] and [l] have been deleted from the actual pronunciation, and are given a small set of frames who are not corresponding to the units [r] or [l] as the canonic transcription says. The training of these units will be incorrect from the point of view of the baseform units; but, in the test signal (bottom left signal), the process of reduction was the same, so the matching between training and testing utterances in the speaker will allow this adaptation to achieve fine performance.

On the right side, acoustic-lexical adaptation is aware of these phonetic reductions and segments the train signal (upper right signal) according to the real transcription [abo], providing a correct adaptation for all units. In the testing phase, the system is set to decode both transcriptions: baseform [arbol] and learned [abo]. As the test signal (bottom right signal) also contains [abo]; the models trained can fit the test input frames and the decoding is correct for the sequence [abo], which is associated to the word 'árbol'.

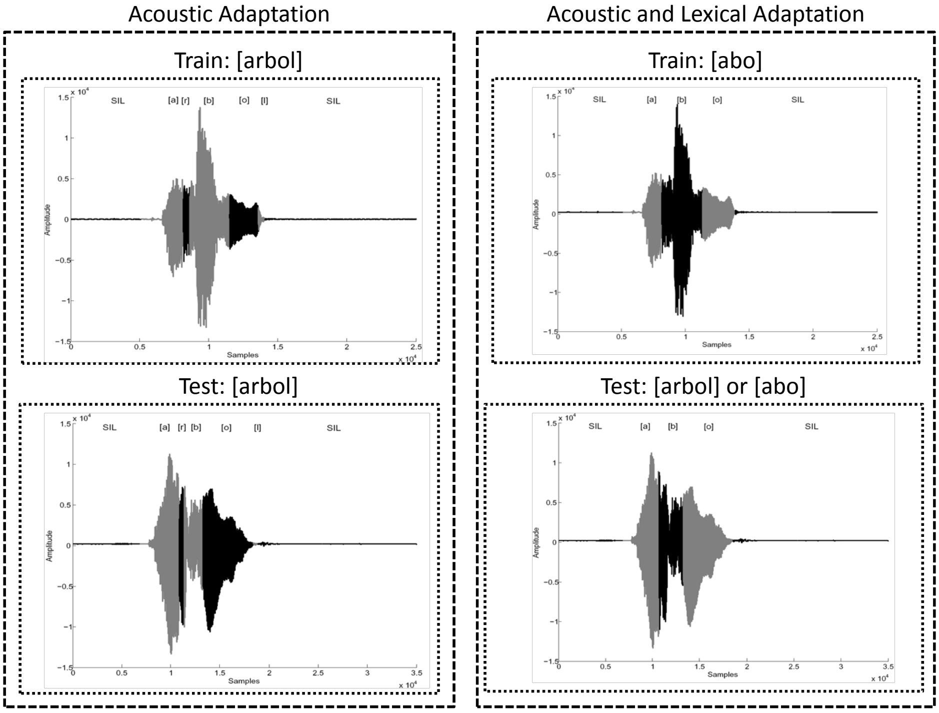


Figure 11.1: Acoustic vs acoustic-lexical personalization in ASR

The case for study in Example 11.3.1 indicates how acoustic and acoustic-lexical adaptations are reaching the same output (the recognition of the second utterance of 'árbol' from the training of the first utterance of 'árbol') by two different ways. In the acoustic approach, the mispronunciations are learned by introducing the variability they provide in the acoustic models; and in the acoustic-lexical approach, the acoustic models are maintained from correct pronunciations and the variability of mispronunciations is introduced in the new lexicon for testing.

11.3.2 Acoustic vs lexical adaptation for PV

The same adaptation techniques used for ASR were tested for the PV task. In this case, the results achieved had different implications than those of ASR: The models trained only with lexically accurate data outperformed the models trained with the baseform transcription in the detection curves and EER in Section 8.4 (15% vs 19%). In this new task, these models are providing such different performances because they are actually doing different things when it comes to the evaluation.

Example 11.3.2 Similarly to Example 11.3.1, this example can show the differences between considering the lexical mistakes of the user in adaptation frameworks for PV with the help of Figure 11.2. This example shows once again the two utterances of the word 'árbol' by Speaker *Spk*05 from the disordered speech corpus, again with deletions in the sounds [r] and [l].

Using the acoustic adaptation based on the baseform transcription, the forced alingment again decodes the phoneme boundaries in the upper left signal with the inaccurate phonemes [r] and [l], that have been deleted from the actual pronunciation, being used to re-train the corresponding units. In this task, the test signal (bottom left signal) has to be evaluated to detect the incorrectness of phonemes [r] and [l], process whose difficulty is a priori increased by the fact that similar inaccuracies have been used in the train phase.

On the right side, the adaptation that is aware of these lexical inaccuracies segments the train signal (upper right signal) according to the real transcription [abo], providing a correct adaptation for all units. In the testing phase, the system is set to detect the mispronunciations of [r] and [l] starting from the baseform transcription [arbol] of the test signal (bottom right signal). In this case, the models for [r] and [l] do not match the incorrect segments shown on the signal and the score can be presumed to be lower, improving the detection accuracy.

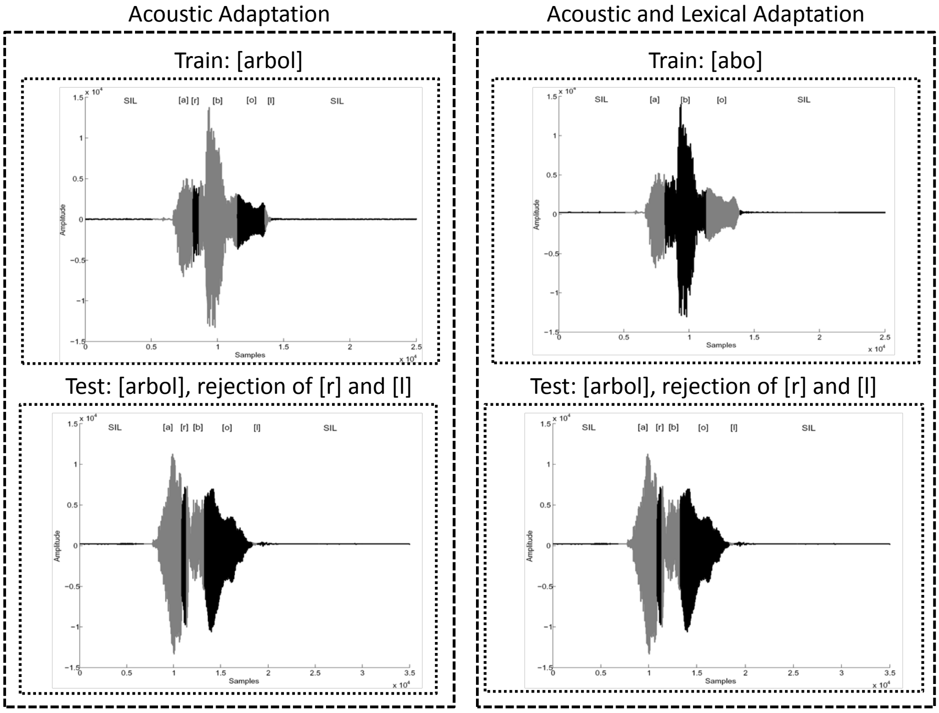


Figure 11.2: Acoustic vs acoustic-lexical personalization in PV

This is giving the clue that the model which only adapts to lexically correct data is more accurate and precise from the acoustic point of view, and that could in certain experiments of ASR provide some gains who were not observed in the current task and domain.

One of the discussions that this thesis has left unclosed is whether these lexical inaccuracies might, at a given moment, produce a degradation also in the performance of the SD acoustic models for ASR as it has been seen in PV. It is believed, after all the experiments made, that this may happen; so the proposal of this thesis is to separate in all cases the influences of the acoustic and lexical degradation in the speech to avoid these possible side effects in the retraining of acoustic models.

11.3.3 Issues on unsupervised adaptation

Finally, the last task in the thesis has been the translation of the different adaptive techniques to an unsupervised framework, where no preliminary information was known about the input signals for adaptation.

The major difficulty in this task has been the low rates in recognition achieved by the baseline systems, due to the heavy disorders of the speakers, which limited the possibilities of providing a reliable estimation of the transcriptions of the input data to the adaptation algorithms. These limitations, that supposed an average improvement of only 14-22% over the TD models, could not be avoided by the use of a confidence measuring system to reject those phonemes which did not match the real pronunciation of the speaker.

The limited size of the test data has been in this case relevant to difficult this work in unsupervised systems, as the models were suffering problems of undertraining as the amount of speech that was accepted by the confidence measure became smaller. This was ratified by the analyses on the influence of the amount of adaptation data in Chapter 7, which showed how the adaptation systems where working in a point still sensitive to the amount of data (only a maximum of 3 sessions for adaptation from each speaker).

However, the results in the proposed on-line personalization framework were promising, as it achieved up to 12% in relative improvement to the baseline, even considering the sparsity of the data which was faced during the successive operations in which the system was working. It is believed that more iterations of the on-line personalization system introduced in Chapter 9 with more data on each iteration might end up obtaining larger improvements.

Several proposals for fixing the problem of data sparsity have been proposed in the literature, including MLLR approaches [Siohan et al., 2000, Chen et al., 2000, Tang and Rose, 2007]. As MLLR, most of them create regression classes which incorporate data from different units to increase the data which is actually used to retrain a given unit. However, it was decided to not use these techniques massively during this thesis, because in all cases, it was intended to keep control on what was being adapted for each unit. With these techniques, not only lexically correct and incorrect versions of a given unit were used to retrain that unit, but also correct and incorrect pronunciations of similar units, increasing the variability of each unit. These approaches would be especially problematic when retraining models for PV, because it would diminish their ability to distinguish correct from incorrect units. For all those reasons, MAP has been extensively used during this thesis, despite the problematic of data sparsity, because it allowed to know at every time what was being fed for the retraining in all cases.

In the end, when a real ASR system is to be deployed as part of an aiding device for a real user, some other factors apart from the purely technological have to be considered to achieve a success in the task. They are mostly two:

- Decide whether a user is really a tentative user of speech technology or not. A good example of this is Speaker *Spk*13, who never achieved a WER lower than 58% (Table 7.1 in Section 7.1). With this rates, no speech-based system is going to work reliable for her, so other possible interaction methods should be designed for her. This decision can be made upon her rates of mispronunciation, where she can hardly pronounce correctly a 45% of the phonemes. In a similar case is *Spk*05, although his results are still improvable, as he shows improvements with the different techniques evaluated. The decision also has to be influenced by the physical characteristics of the user, as the speech process sometimes can be more tiring and demanding than other physical movements or control.

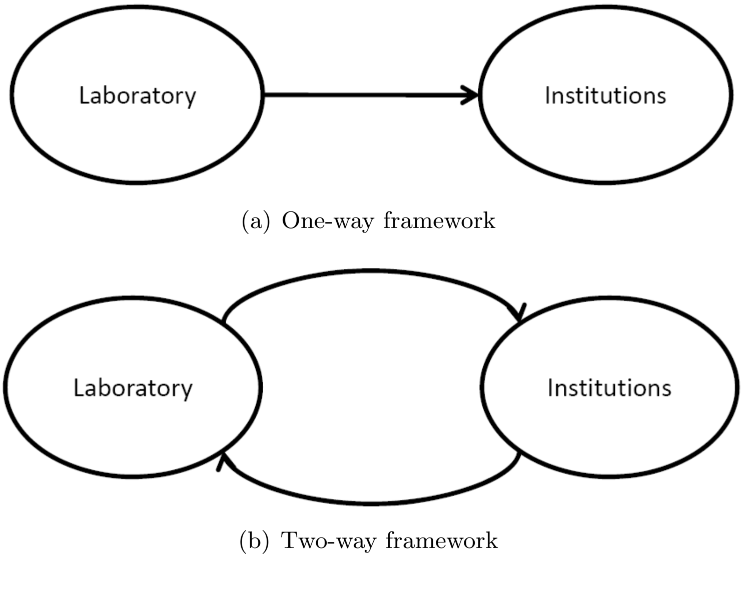
- A correct selection of the vocabulary, as the results achieved in this thesis have been obtained with a small-medium vocabulary of 57 words with special phonological difficulties. In a real situation, smaller vocabularies which are especially chosen for each speaker according to their characteristics would boost the results for all of them.

11.4 Translation of the Research Work into the Deployment of Real Applications

Finally, the last element for discussion are the implications that the work carried out within this thesis has had and will have for the development of applications in the real world that can effectively provide a better quality of life for a part of the handicapped community.

Regarding how this thesis has influenced the development of tools during these latest years, it was seen in Chapter 10 that the knowledge achieved in ASR of disordered speech and PV of lexical mispronunciations has had a capital role in the development of the CAPT tools in "Comunica": “Vocaliza” and “VocalizaL2”. Further theoretical research [Rodríguez and Lleida, 2009] has also been relevant for the development of other tools like “PreLingua”; showing how, in many cases, it is possible to translate the work and experiments in the laboratory to produce an effect in the world outside the lab.

The experience gained during these years has shown that the effort in creating these novel tools can not be fully effective if the work in the laboratories and in the outside world were to be dissociated. Although theoretically, the process flows from the laboratory to the real world as in Figure 11.3(a); the feedback that educators, therapists and assistants can provide in the early stages of development supposes a very necessary information that can lead to changes in the way in which scientist and technicians can affront their research as in Figure 11.3(b). The development of “Comunica” would have never been the same if the opinions of the therapists had not been considered.



(a) One-way framework

(b) Two-way framework

Figure 11.3: Frameworks of collaboration academia-institutions

Another key point is to find a situation of agreement between the objectives of the two elements implicated in this work (academia/industry and assistance institutions). If laboratories and industry try to fulfill only their research objectives and translate directly that knowledge to the institutions like in Figure 11.3(a), the work will never be so fruitful as if there is a two-channel communication as in Figure 11.3(b) where the institutions can correct the ideas of the academia according to their own needs and requirements, as long as they also help to fulfill the research objectives of the academia.

11.4.1 The role of public agencies

One of the major points of discussion during these years has been the special implication that public institutions need to put in this effort to advance further in the accomplishment of the objectives. As it was well pointed out in [Cucchiarini et al., 2008a], governments and public agencies have to play the biggest role in the development of these aids, because following the criteria of industrial benefits, they are far away from being profitable unless a high price is set. This is due to the fact that the handicapped population represents a small percentage of the global population and furtherer due to the high variability of each individual's handicap and the big requirements for personal adaptation that each aiding tool requires for its final deployment.

Figure 11.4 pretends to be a representation of what it is understood from the experience of these years as the optimum framework for the development of aiding tools for the handicapped: The public agencies are the umbrella that covers the joint work that academia, industry and assistance institutions face during the complex process of development of these tools.

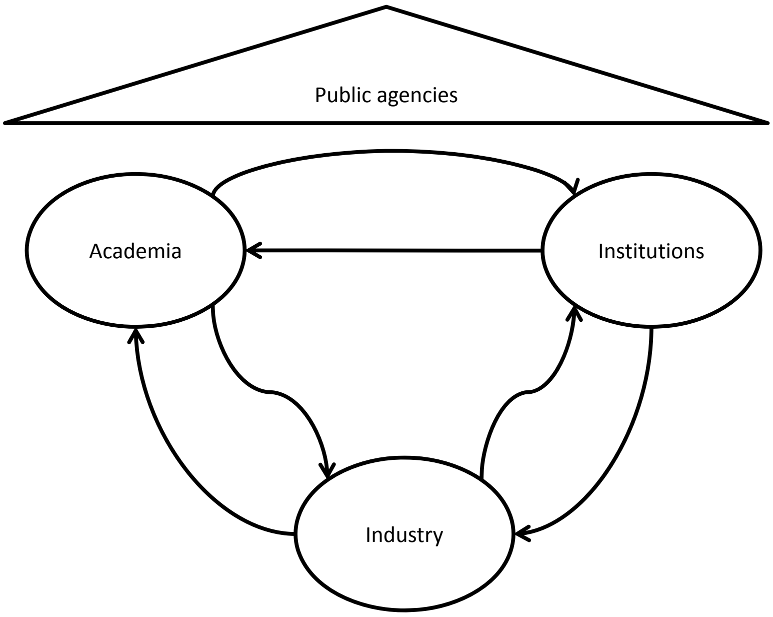


Figure 11.4: Optimum framework of collaboration

With this proposed framework, public agencies can stimulate the apparition of novel research efforts in this area with the cooperation of the different institutions who can benefit from that work. When sufficient experience has been gained, it is possible then to move to the industry to propose the possibilities of transmitting the acquired knowledge to the society via novel developments of software or hardware. The economical difficulties that this might suppose could be covered then by the public agencies again, as their way of investing in policies for the improvement of the quality of life of their disabled individuals.

A case in this direction which was witnessed during the work in this thesis has been the development of a set of more than 3500 pictographs in the Portal Aragonés de la Comunicación Aumentativa y Alternativa (Aragonese Portal for the Augmentative and Alternative

Communication) (ARASAAC) portal[1]. Pictographs are a basic tool in AAC elements [Beukelman and Mirenda, 1998], as they allow for non-verbal communication to individuals with language disabilities. The creation of a set of these pictographs can be time demanding for a graphic designer, and so, commercial pictographs for this purpose tend to be expensive in many cases.

In the case of ARASAAC, a public agency like the Centro Aragonés de Tecnologías de la Educación (Aragonese Center for Technologies in Education) (CATEDU) appointed part of the staff of the CPEE "Alborada" to work in collaboration with a graphic designer in the creation of this novel set of pictographs which could be distributed under a Creative Commons (CC) license. The CATEDU provided the fundings to pay the extensive work of the designer, who followed the instructions of the specialists in the CPEE "Alborada" to create the pictographs according to the needs of the special users of these AAC elements.

Nowadays, ARASAAC is a reference in the world of AAC and has become a major resource for all elements requiring AAC elements, including the CASLT tools in "Comunica". "Vocaliza", for instance, includes in its download a set of 500 words ready to be trained within the application which incorporate their pictographs from the ARASAAC resources.

11.4.2 The experience of "Comunica"

"Comunica" has tried to be an effective way of fulfilling the development of state-of-the-art tools for CASLT while achieving the objectives which are usually the hardest to reach within the academia: the successful collaboration and feedback with institutions and users. For this reason, a lot of effort has been put in the distribution of the tools and the communication with the end users. This relationship has been based in the website of "Comunica"[2] and in all the possibilities that this offers to developers and users.

The different elements of interaction have been many through the site, and a small summary of them since the launching of the site in February 2008 is in Table 11.1. The most remarkable outcome is the number of registered users (around 5000) from all over the world; this is indicating approximately the number of potential users of the tools, as the simple registry (providing an e-mail address) is required for downloading the tools.

Table 11.1: "Comunica" statistics (as recorded by October 23rd, 2009)

Element	Data
Website date of start	February 1st, 2008
Registered users	5,201
Answers in the Frequently Asked Questions	30
Testimonials	11
Mails exchanged	216
Bulletins sent	9
Reviews collected	34

This direct communication with the end users fructified in many improvements over the tools, fixing bugs or adding new capabilities, and in the development of the latest tools in "Comunica": "ReFoCas" and "ReFoGal", presented in Section 10.2.4. The idea for these tools came directly from a group of users in different educative centers in Spain and; after a discussion with the development team, it lead to this novel tool for the easing of the process of evaluating phonological registers in young students.

[1] http://www.catedu.es/arasaac

[2] http://www.vocaliza.es

More statistics have been collected from the website; for instance, the number of unique visitors since the 22nd week of 2008 is shown on Figure 11.5. Average number of visitors per week is 346 with the larger peaks in Week 12 of 2009 (895 visitors) and Week 33 of 2009 (1464 visitors) coinciding with the apparition of "Comunica" in Aragón Radio, the radio network of the Corporación Aragonesa de Radio y Televisión (Aragonese Corporation of Radio and Television) (CARTV), and the presentation in the 8th Iberoamerican Conference in Informatics and Special Education [Rodríguez et al., 2009], respectively.

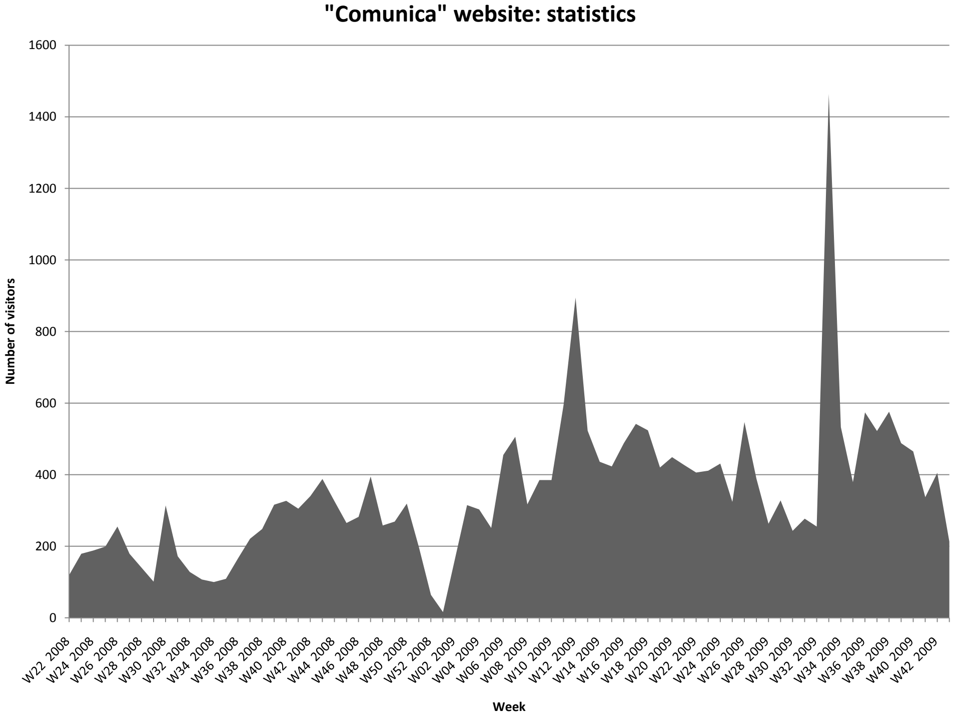


Figure 11.5: Visitors to the "Comunica" website (as by the 23rd of September of 2009)

Other notable appearances in the media, with the help it supposed to the visibility of the tools, have been the presentation of "VozClick" in Huesca (Spain) through different notes in local newspapers and radio networks. The launching of "Vocaliza" also supposed its appearance in the television network of the CARTV and in local newspaper as the author of its work, Carlos Vaquero, was given different awards for his undergrad thesis, including the special award of the Capítulo Español de la Sociedad de la Educación del IEEE (Spanish Area of the IEEE Education Society) (CESEI).

The wide dissemination of "Comunica" has supposed that many of the current users are in Latin America. These tools have had an special impact in the whole community of therapists in these countries, where several of the therapists have contacted the developers to congratulate their work and propose new ideas and visions.

As it was stated in Chapter 1, all this research work only makes sense if it is intended to be useful for a potential user. The research work, hence, carried out in this thesis and in the whole VIVOLAB group has been as successful for its research results which have lead to many different scientific publications as for the gates that it has opened to the communication with all these individuals, potential users of this technology.

Chapter 12

Conclusion

All the above shown accusations against me are based on truth

Josef Švejk

-Jaroslav Hašek, *The Good Soldier Švejk: and His Fortunes in the World War*

The last Chapter of this thesis is oriented to be a short summary of the work done during these years, while pointing out the future lines of research that this thesis has left opened. This whole work has to be framed within the ongoing interest of the community of speech researchers for the development of algorithms and tools for improving the quality of life of handicapped individuals; and, more precisely, within the effort started in parallel to this thesis in collaboration with different institutions.

Section 12.1 will summarize the key of points of the thesis, providing a quick approach to all the main points of interest, while Section 12.2 will point out how this work has fulfilled or not the objectives set at the beginning of the thesis. Finally, Section 12.3 will indicate which are thought to be the most interesting areas for future research opened by this thesis.

12.1 Brief Summary of the Thesis Work

Once the end of this thesis has been reached, it is necessary to provide a comprehensive view that covers all the work carried out in it, summarizing all its relevant achievements. This way, all the major results through the thesis can be evaluated at a glimpse:

- First, a fully functional corpus for research has been acquired, gathering speech from young speakers suffering a wide range of speech disorders; altogether with a set of age-matched peers. The whole corpus, with more than 2 hours of disordered speech and more than 8 hours of speech from unimpaired individuals contains, hence, sufficient amount of data to be useful and significant for speech research. This corpus has been post-processed to obtain a perceptual labeling of the lexical mispronunciations made by the impaired speakers with a high consistency between the humans, making it reliable to serve as ground truth for the development of automated mispronunciation detection algorithms. Both the corpus and the human labels are freely available for research purposes to any group interested in it, as long as the requirements of privacy of the speakers are maintained. The sharing of corpora and resources in this area of research is expected to boost the interest of the different research groups and individuals.

- Acoustically, a certain decrease in the quality properties of this speech has been observed and measured. Only vowels underwent this analysis, as consonants would require an articulatory analysis which was not possible only from the acoustic signals. The major effects of degradation appeared in the formant map of the speakers, whose separability was lowered among main vowels like /a/, /e/ and /o/ around 20% or more; and also in the durational cue of these vowels, which become extremely variant in their production, leading to an unsteady production of speech. Furthermore, the ability of these speakers to produce a distinction in terms of intensity between stressed and unstressed vowels is totally reduced; indicating the existence of phonatory inabilities in their speech. Anyways, in general terms the speakers did seem to control suprasegmental features consistently.

- It has been from the lexical point of view where the disorders of the speakers have offered more interpretations and possibilities for analysis. The development impairments of most of the speakers have supposed a delay in their language abilities which has made their speech closer to small preliterate children, as proved by the comparison of the patterns of mispronunciations between both types of groups (disordered speech in this thesis and children in [Bosch-Galcerán, 2004]). The lexical inaccuracies of this group of speakers has been measured as especially large in vibrants and laterals (up to 50% of mispronunciations); but the main effect appeared in the study of the influence of the syllabic context; where it was seen that diphthongs, codas and consonant clusters provoked a significant increase in the mispronunciations of the speakers. This study should be complemented with previous knowledge in the areas of neurolinguistics and psycholinguistics to better understand whether it is a pure delay in language acquisition or a problem in the general capabilities of the speakers.

- The influence of these separated effects (acoustic and lexical disorders) has lead to a dramatic loss of performance in the baseline ASR system (from 2-3% of WER with the unimpaired speakers to 40% with their impaired peers). The use of speaker adaptation, with acoustic and lexical approaches, has been evaluated leading to very promising results (nearly 50% of improvement with SD acoustic models and SD lexicon). One of the more remarkable results has been the understanding of how acoustic and lexical adaptation can, when merged properly, overcome the initial limitations of each other. Lexically, only data-driven methods were seen as possible for these speakers, as each speaker was following an independent way of producing mispronunciations from the whole set of lexical variabilities that the impaired speech group was offering.

- The need of an automated detection of mispronunciations has led to the study on how to avoid speaker variability in PV methods. Score normalization and speaker adaptation frameworks have been proposed to work in this task of detecting lexical mistakes according to the labels assigned by the human experts. The results have been particularly promising in this area, leading to very interesting collaborative work [Yin et al., 2008, Yin et al., 2009], and achieving a performance that can be considered as state-of-the-art for the proposed task. The study on how the proposed methods are related and help to reformulate traditional techniques like GOP as an score normalization technique has been an interesting result too, as it opened the gate for more research in new combination of techniques. Furthermore, the thesis has studied how adapting the acoustic models used in the confidence measure system can led to an improved performance, but a certain attention has to be paid to the fact that mispronunciations existing in the training data can lead to a loss of accuracy in the models and a loss of performance in the overall PV system.

- Unsupervised personalization systems have been studied and evaluated. Their results have not reached the results that supervised techniques could predict, because there was a heavy effect regarding the low accuracy results on the preliminary recognition phase over the adaptation data. It has been also seen how the sparsity of the available data has limited the possibilities of the proposed unsupervised techniques, because reducing the number of phonemes used in adaptation via the confidence measure system lead to a non-sufficient amount of data in the adaptation framework. Finally, a fully on-line system has been proposed based on ASR, PV as confidence measure and MAP adaptation to perform personalization of the systems as the user interacts with it in several steps. In the end, different approaches have been evaluated, but the need of further work has been required to improve and support the work carried out in this area in the thesis.

- Finally, all the experimental research in this thesis has gone parallel to the work for the development of tools for aiding impairment. The collaboration with different institutions for education and for assistance of the handicapped has been fruitful and the knowledge transferred from them has supposed the first stages of development of oral control devices and a big effort with "Comunica" and its tools for CASLT in different linguistic levels. The discussion in Section 11.4 has presented the different elements that can report "Comunica" as a case of success in the creation of novel semi-automated tools for speech therapy.

12.2 Fulfillment of Goals and Objectives

Reviewing the set of objectives proposed for this thesis in Section 1.4, they were divided in research objectives and development objectives. The evaluation of the grade of completion of these objectives will be made, hence, following the same criteria of separation.

12.2.1 Research objectives

The three set of research objectives have been accomplished in different degrees:

- The acquisition of the corpus was the first goal of the work; it was a prerequisite for the completion of this thesis, as the evaluation of the developed algorithms needed a benchmark for testing; and unfortunately, no one of the available resources studied could completely fulfilled them. This objective has been fully accomplished with the acquisition of the "Alborada-I3A" corpus: This corpus contains sufficient data for obtaining significant enough results as it was the outcome of this thesis. Furthermore, the acoustic and lexical analyses of this corpus, which were not set initially as objectives, have been carried out extensively and have had a major relevance in the work as it may help to understand the sources of errors in the speech technology-based systems.

- Regarding all the proposed objectives in adaptation and personalization techniques, the goal of obtaining knowledge in the influence of acoustic and lexical speaker adaptation in the performance of ASR systems has been achieved to a deep level. The proposals of different frameworks for acoustic adaptation have been useful to understand how the diverse transcriptions affect the results in the recognition system. Furthermore, the data-driven lexical adaptation proposed in this work, based on APD, has shown very interesting results for adapting to the special lexical variability of these users; with the interaction between acoustic and lexical adaptation frameworks as a very interesting outcome of this thesis. Unfortunately, the proposals for unsupervised adaptation still have not achieved a results close to the better possible scenario with supervised data. Although these unsupervised results significantly overcome the baseline system, this is the point of this thesis where more work should be set in the future, specially in the development of real-world applications.

- Finally, the objective of deriving an automated method for the detection of mispronunciations by speakers with special lexicon needs has been fully achieved with the proposal made in this thesis. The parallelism with the speaker verification task has shown to fit perfectly the PV task and its special characteristics, proposing techniques for score normalization and speaker adaptation which have been adapted successfully to the task. In the end, the results have been satisfactory and ready to be used into CAPT tools like "VocalizaL2".

12.2.2 Development objectives

Regarding the development of real-world applications for handicapped aid, which was set as another different group of objectives at the beginning of the thesis, the grade of completion of them has been diverse too during these years.

- The possible development of oral command control interfaces for the handicapped has been fulfilled only in preliminary attempts like the oral control of the robotized wheelchair or the VAD control of mouse and other computer elements. The late development of the unsupervised adaptation algorithms has not allowed for their introduction in any application which can be ready at the date of today. However, all the acquired knowledge will be useful for the possible development of oral-controlled interaction elements.

- On the contrary, the development of CALL tools, with special interest for the handicapped has fulfilled and even overpassed the expectancies. All the work in "Comunica", closely related to this thesis, has fructified in the development of several tools ("PreLingua", "Vocaliza", "Cuéntame", "VocalizaL2", "ReFoCas") which make and extensive use of speech technologies to provide language teaching abilities to handicapped or foreign users: ASR, speaker adaptation or PV. The tools have been evaluated in several environments and a big effort has been put in the development of appropriate interaction interfaces suitable for the target users via AAC elements. As it was shown in the discussion in Chapter 11, the communication and feedback from users through the website has been massive and really encouraging about the actual state of development and future possibilities in this area.

12.3 Future Lines of Work

The future lines of research that this thesis has opened are mainly the completion of some of the subjects that have been covered but not completely fulfilled in the work these years. Also, the discussion of the results of this thesis has pointed out some new proposals for work in the short-term future.

- Although some studies have been carried out, it has been seen in the review of the objectives in previous Section that unsupervised techniques for adaptation have been the main objective that this thesis could not fulfill completely. The approaches evaluated in this thesis have been successful in providing a certain improvement over the baseline systems, but no improvement could actually be obtained over the initial framework for unsupervised adaptation. The natural limitations of the confidence measure proposed here, although a 15% WER was achieved, and the problem of data sparsity have reduced the possibilities of the proposed approaches. Once this thesis has help to understand how acoustic and lexical variants influence the recognition of this speech, future works might use several MLLR-based techniques proposed for fast adaptation which were not evaluated in this thesis to maintain control of the adaptation data used to retrain each unit. With these techniques solving the problem of data sparsity, performance might be boosted easily in the ASR task.

- Regarding PV techniques, during this thesis it was shown how distinguishing between lexically correct and incorrect units in the adaptation phase helped to provide better evaluation performance. One of the future works has to be directed to achieve this improved performance in an unsupervised way; this is, without the a priori knowledge provided by the human experts. Other possibilities might include the use of garbage models which gather the knowledge on lexically incorrect units to be used in UV-like scheme like in [Lleida and Rose, 2000]. This extra anti-units or garbage units might be useful to provide an enhanced separability between correct and incorrect phonemes in a utterance.

- An interesting possibility would be the evaluation of the proposed frameworks for speaker adaptation and confidence measuring in different corpora that might incorporate a different relationship between acoustic and lexical disorders in their speech. Works with dysarthric speech like the UADatabase [Kim et al., 2008] might be, among others, an interesting testbench for these new set of studies and evaluation. The comparison between both types of speech (dysarthric speech vs acquisition delayed speech) would provide a very interesting discussion on the differences in how to deal with these two both types of disordered speech, as the use of acoustic and lexical adaptation might be different in new domains.

- New possibilities for developing oral-based command control tools have arisen at the end of this thesis. The possibility of including an adaptable ASR system in the "Ratón Virtual" [Bergua, 2005] framework is now possible with the new driver control system included in this device [González, 2009]. The main work in this case should be directed to include personalization techniques that allowed every user of the device to interact with an enhanced performance with the system. The close framework of collaboration with institutions like the CPEE "Alborada" and ASPACE-Huesca has to encourage this work, as they count with tentative users of this technology who might serve as test users of the novel tools.

- Involving "Comunica", it is expected to go on with the improvement of the existing tools, once all facets of language have been covered with the present tools. A major effort has to be carried out to make the tools more accessible to all the community of speech therapists or handicapped individuals, including the development of the tools for other Operative Systems like MacOS or Linux or the development of Internet-based versions of the tools. The next step, once all the current tools have faced their current improvements, are the highest levels of language to be worked with reading tutors or text-based tools which aim to improve the grammatical abilities of the students, as well as their vocabulary and syntax. It is expected that more research groups can get involved in this framework to collaborate in these future lines of work or to handle possible translations of the tools to different languages.

Appendix A

Review of Spanish Phonetics

España y yo somos así, señora

Eduardo Marquina, *En Flandes se ha puesto el sol*

This Appendix aims to provide to those readers that might be unaccustomed to the Spanish phonetics with a brief description of the sounds and allophones of the Spanish language. Although the phonetic system of Spanish can be seen as much simpler than other languages of its environment which contain a larger number of phonemes and sounds (for instance, English); the massive use of phonetic references and instances in the thesis makes necessary this Appendix.

The Sections of this Appendix provide the whole list of Spanish phonemes, their corresponding allophones and their SAMPA and IPA transcriptions in Section A.1 and describe the most usual phonetic reductions in peninsular Spanish in Section A.2.

A.1 Set of Spanish Phonemes and Allophones

Castilian Spanish is traditionally described in 24 phonemes and a set of 51 sounds or allophones, where allophones are realizations of a same phoneme with different articulatory features. To ease the reading of this Appendix, phonemes will be gathered according to the usual distinction of manner of articulation.

A.1.1 Vowels

Spanish contains 5 vowels (/a/, /e/, /i/, /o/ and /u/). Their description is as follows (summarized in Table A.1):

/a/

Three allophones define the pronunciation of the vowel /a/.

- [a] is the central low vowel. It appears in all contexts for /a/ except for the other two allophones.
- [ɑ] is the posterior low vowel allophone appearing in the following contexts:
  - Before a vowel /u/, either if they form a diphthong or a hiatus.
  - Before the vowel /o/.

  - Before the consonant /x/.
  - In all syllables finishing in consonants.

- [ã] is the nasalized central vowel appearing in between nasal consonants or after a nasal consonant at the end of syllable.

/e/

Three allophones define the pronunciation of the vowel /e/.

- [e] is the front medium vowel. It appears in all contexts for /e/ except for the other allophones.
- [ɛ] is the front low vowel allophone appearing the following contexts:
  - In contact with [r] except when the syllable is finished in any of these consonants ([d], [m], [n], [s] or [T]).
  - Before the consonant /x/.
  - When forming a closing diphthong with /i/.
- [ẽ] is the nasalized central vowel appearing in between nasal consonants or after a nasal consonant at the end of syllable.

/i/

Six allophones define the pronunciation of the vowel /i/.

- [i] is the front high vowel. It appears in all contexts for /i/ except for the other allophones.
- [i̞] is the front medium-high allophone appearing in syllable finished in consonant, before or after [r] and before [x].
- [ĩ] is the nasalized front high vowel appearing in between nasal consonants or after a nasal consonant at the end of syllable.
- [ĩ̞] is the nasalized front medium-high vowel appearing in between nasal consonants or after a nasal consonant at the end of syllable.
- [j] is the front glide appearing before or after another vowel creating a diphthong or tripthong.
- [j̃] is the nasalized front glide appearing in between nasal consonants or after a nasal consonant at the end of syllable.

/o/

Four allophones define the pronunciation of the vowel /o/.

- [o] is the labialised posterior medium vowel. It appears in all contexts for /o/ except for the other allophones.
- [ɔ] is the labialised posterior medium-low allophone appearing in the following contexts:
  - In contact with [r].
  - Before the consonant [x].
  - In closing diphthong with i.

  - In a syllable finished by any consonant.
  - In a tonic position between a preceding [a] and a following [r] or [l].

- [õ] is the nasalized labialised posterior medium allophone appearing in between nasal consonants or after a nasal consonant at the end of syllable.

- [ɔ̃] is the nasalized labialised posterior medium-low allophone appearing in between nasal consonants or after a nasal consonant at the end of syllable.

Table A.1: Spanish Phonemes (Vowels)

Phonemes		Allophones		
IPA	SAMPA	IPA	SAMPA	Examples
/a/	/a/	[a]	[a]	c***a***sa
		[ɑ]	[A]	j***a***ula; t***a***o; ***a***jo; ***a***rbol
		[ã]	[a~]	m***a***no; campan***a***
/e/	/e/	[e]	[e]	caram***e***lo; car***e***ncia; ***e***scoba
		[ɛ]	[E]	car***e***ta; t***e***jado; p***e***ine; pu***e***rta
		[ẽ]	[e~]	m***e***nos; ***e***mpezar
/i/	/i/	[i]	[i]	mar***i***posa
		[i̞]	[I]	lap***i***z; perr***i***to; ***i***rritar; h***i***jo
		[ĩ]	[i~]	n***i***ño
		[ĩ̞]	[I~]	m***i***ntió; R***i***n; ***i***ndio
		[j]	[j]	ind***i***o; ovo***i***de
		[j̃]	[j~]	Tan***i***a
/o/	/o/	[o]	[o]	gl***o***b***o***
		[ɔ]	[O]	perr***o***; ***o***jo; h***o***y; arb***o***l; mata***o***r
		[õ]	[o~]	m***o***no; ***o***ntología
		[ɔ̃]	[O~]	r***o***n; m***o***nte
/u/	/u/	[u]	[u]	d***u***cha
		[u̞]	[U]	r***u***ta; ag***u***ja; t***u***rno
		[ũ]	[u~]	m***u***nición
		[ũ̞]	[U~]	m***u***ndo; r***u***mba; ***u***ntar
		[w]	[w]	p***u***eblo
		[w̃]	[w~]	ten***u***e

/u/

Six allophones define the pronunciation of the vowel /u/.

- [u] is the labialised posterior high vowel. It appears in all contexts for /u/ except for the other allophones.

- [u̞] is the labialised posterior medium-high allophone appearing in syllable finished in consonant, before or after [r] and before [x].

- [ũ] is the nasalized labialised posterior high allophone appearing in between nasal consonants or after a nasal consonant at the end of syllable.

- [ũ̞] is the nasalized labialised posterior medium-high allophone appearing in between nasal consonants or after a nasal consonant at the end of syllable.

- [w] is the posterior glide appearing before or after another vowel creating a dypthong.
- [w̃] is the nasalized posterior glide appearing in between nasal consonants or after a nasal consonant at the end of syllable.

A.1.2 Plosive Consonants

There are 6 plosive consonants in Spanish (summarized in Table A.2):

/p/

One allophone defines the pronunciation of the plosive consonant /p/.

- [p] is the bilabial unvoiced plosive consonant. It appears in all contexts.

/b/

Two allophones define the pronunciation of the plosive consonant /b/.

- [b] is the bilabial voiced plosive consonant. It appears in the following contexts:
  - In the beginning of word after a pause.
  - In the inning of a group, after a nasal.
- [β̝] is the bilabial voiced approximant consonant, appearing in all cases except in the ones defined for [b].

/t/

Two allophones define the pronunciation of the plosive consonant /t/.

- [t̪] is the dento-alveolar unvoiced plosive consonant. It appears in all cases except in the defined for other allophones.
- [t̪̟] is a interdental voiced plosive consonant, appearing after a [θ].

/d/

Two allophones define the pronunciation of the plosive consonant /d/.

- [d̪] is the dento-alveolar voiced plosive consonant. It appears in the following contexts:
  - In the beginning of word after a pause.
  - Before or after /n/ or /l/.
- [đ̝] is the interdental voiced plosive consonant, appearing in all cases except the ones defined for the previous allophone.

/k/

One allophone defines the pronunciation of the plosive consonant /k/.

- [k] is the velar unvoiced plosive consonant. It appears in all cases.

/g/

Two allophones define the pronunciation of the plosive consonant /g/.

- [g] is the velar voiced plosive consonant. It appears in the following contexts:
  - In the beginning of word after a pause.
  - Before or after a nasal.
- [ɣ̞] is the velar voiced approximant consonant, appearing in all cases except the ones defined for the previous allophone.

/ʝ̞/

Two allophones define the pronunciation of the approximant consonant /ʝ̞/.

- [ʝ̞] is the palatal voiced approximant consonant. It appears in all cases except the ones defined for the other allophone.
- [ɟ͡ʝ] is the palatal voiced africate consonant. It appears at the beginning of syllable, immediately after /n/ or /l/.

Table A.2: Spanish Phonemes (Plosives)

Phonemes		Allophones		
IPA	SAMPA	IPA	SAMPA	Examples
/p/	/p/	[p]	[p]	***p***ueblo
/b/	/b/	[b]	[b]	***b***oca; am***b***os
		[β̞]	[B]	glo***b***o
/t/	/t/	[t̪]	[t]	***t***ren
		[t̪̟]	[ts]	ac***t***uar
/d/	/d/	[d̪]	[d]	***d***edo; in***d***io; mol***d***e
		[đ̞]	[D]	de***d***o
/k/	/k/	[k]	[k]	***c***asa; ***qu***eso; ***k***ilo
/g/	/g/	[g]	[g]	***g***orro; in***g***lés
		[ɣ̞]	[G]	a***g***ua
/ʝ̞/	/j\/; /jj/	[ʝ̞]	[j\]; [jj]	pla***y***a
		[ɟ͡ʝ]	[J\]; [JJ]	cón***y***ge

A.1.3 Nasal Consonants

There are 3 nasal consonants in Spanish (summarized in Table A.3):

/m/

One allophone defines the pronunciation of the nasal consonant /m/.

- [m] is the bilabial voiced nasal consonant. It appears in all cases of ‘m’ and with ‘n’ before a bilabial consonant.

/n/

One allophone defines the pronunciation of the nasal consonant /n/.

- [n] is the alveolar voiced nasal plosive consonant. It appears in all cases.

/ɲ/

One allophone defines the pronunciation of the nasal consonant /ɲ/.

- [ɲ] is the palatal voiced nasal plosive consonant. It appears in all cases of 'ñ' and with 'n' before a palatal.

Table A.3: Spanish Phonemes (Nasals)

Phonemes		Allophones		
IPA	SAMPA	IPA	SAMPA	Examples
/m/	/m/	[m]	[m]	***m***oto; e***n***vio
/n/	/n/	[n]	[n]	***n***iño
/ɲ/	/J/	[ɲ]	[J]	ni***ñ***o; a***n***cho

A.1.4 Fricative Consonants

There are 5 fricative consonants in Spanish (summarized in Table A.4):

/f/

One allophone defines the pronunciation of the fricative consonant /f/.

- [f] is the labiodental unvoiced fricative consonant. It appears in all cases.

/θ/

Two allophones define the pronunciation of the fricative consonant /θ/.

- [θ] is the interdental unvoiced fricative consonant. It appears in all cases, except in the ones defined for other allophones.

- [θ̬] is the interdental voicing unvoiced fricative consonant. It appears at the end of a syllable in contact with a voiced consonant.

/s/

Two allophones define the pronunciation of the fricative consonant /s/.

- [s] is the alveolar unvoiced fricative consonant. It appears in all cases, except in the ones defined for other allophones.

- [s̬] is the interdental voicing unvoiced fricative consonant. It appears at the end of a syllable before another consonant.

/t͡ʃ/

One allophone describes the pronunciation of the fricative consonant /t͡ʃ/.

- [t͡ʃ] is the prepalatal unvoiced africate consonant. It appears in all cases.

/x/

One allophone defines the pronunciation of the fricative consonant /x/.

- [x] is the velar unvoiced fricative consonant. It appears in all cases.

Table A.4: Spanish Phonemes (Fricatives)

Phonemes		Allophones		
IPA	SAMPA	IPA	SAMPA	Examples
/f/	/f/	[f]	[f]	***f***uma
/θ/	/T/	[θ]	[T]	ca***z***a; ca***c***ería
		[θ̬]	[T_<]	ha***z***me
/s/	/s/	[n]	[n]	***s***illa
		[s̬]	[z]	a***s***ma
/t͡ʃ/	/tS/	[t͡ʃ]	[tS]	du***ch***a
/x/	/x/	[x]	[x]	o***j***o

A.1.5 Lateral Consonants

There are 2 lateral consonants in Spanish (summarized in Table A.5):

Table A.5: Spanish Phonemes (Laterals)

Phonemes		Allophones		
IPA	SAMPA	IPA	SAMPA	Examples
/l/	/l/	[l]	[l]	***l***una, carame***l***o, a***l***mena
		[l̟]	[l_A]	a***l***zar
		[l̪]	[l_d]	al***t***o; mo***l***de
		[l̠]	[l_j]	co***l***cha
/ʎ/	/L/	[ʎ]	[L]	***ll***ave

/l/

Four allophones define the pronunciation of the lateral consonant /l/.

- [l] is the alveolar voiced lateral consonant. It appears in all cases except the defined for the other allophones.
- [l̟] is the interdental voiced lateral consonant. It appears before [θ].
- [l̪] is the dental voiced lateral consonant. It appears before [t] and [d].
- [l̠] is the palatalized voiced lateral consonant. It appears before [tS], [JJ], [L] and [J].

/ʎ/

One allophone defines the pronunciation of the lateral consonant /ʎ/.

- [ʎ] is the palatal lateral consonant. It appears in all cases.

A.1.6 Taps and Flaps

There are 2 taps in Spanish (summarized in Table A.6):

/ɾ/

One allophone defines the pronunciation of the tap /ɾ/.

- [ɾ] is the alveolar tap. It appears in all cases.

/r/

One allophone defines the pronunciation of the flap /r/.

- [r] is the alveolar flap. It appears in all cases.

Table A.6: Spanish Phonemes (Taps and flaps)

Phonemes		Allophones		
IPA	SAMPA	IPA	SAMPA	Examples
/ɾ/	/4/; /r/	[ɾ]	[4]; [r]	ca***r***amelo; pue***r***ta; p***r***eso
/r/	/r/; /rr/	[r]	[r]; [rr]	***r***atón; go***rr***o

A.2 Common Reductions: "Yeísmo"

Phonetic reductions are common in all languages as part of the process in which languages, as living beings, are involved in. In the case of peninsular Castilian Spanish, the two most prominent are the 'ceceo' or 'seseo' and the 'yeísmo' or 'lleísmo".

The *'ceceo'* or *'seseo'* is a common reduction in Spanish from the Andalusian region in which the interdental fricative phoneme /θ/ (/T/) is pronounced as the alveolar fricative /s/ ([s]). This reduction is also fully extended in Latin America and is one of the most distinguishable features of the Latin American dialects of Spanish in comparison to the standard peninsular Castilian Spanish. As mentioned before, the 'ceceo' is very extended in the Andalusian region and also affects some parts of the Valencian community and Galicia for transferences with their own regional languages (Valencia catalan and Galician).

The *'yeísmo* or *lleísmo* is a fully extended reduction in peninsular Spanish in which the approximant palatal /j̞/ ([jj]) is pronounced as the palatal lateral /ʎ/ ([L]) [Calero-Vaquera and Calvillo-Jurado, 1992]. The *yeísmo* started mainly in urban areas, but nowadays has largely transferred to the whole Spain (urban or rural) except in old individuals from the rural context, which still keep the distinction. Furthermore, this reduction is also quite usual in Latin America, except in the southern part of the continent (Argentina, Chile or Paraguay), where two different phonemes are still subsisting, although they might have changed some of their features.

There are many more reductions and phonological processes that affect the Spanish language. Many of them define different dialects of Latin American Spanish; while others appear in Spain in certain areas as transference with other peninsular languages. This small Section only intended to show those two reductions affecting peninsular Spanish, not related to other languages, and which may have an impact in the work of this thesis (i.e. *yeísmo*).

Appendix B

Review on Adaptation Algorithms

> I've studied now Philosophy And Jurisprudence, Medicine
> And even, alas! Theology,
> From end to end, with labor keen;
> And here, poor fool! with all my lore I stand, no wiser than before
>
> -Johann Wolfgang von Goethe, *Faust*

This Appendix brings a brief review of the two most popular algorithms for the re-training of acoustic SD models: Maximum A Posteriori (MAP) and Maximum Likelihood Linear Regression (MLLR). As an extensive use of speaker adaptation techniques has been made in this work, this review describes the main features of both algorithms, providing the actual implementation used in the experiments carried out in this thesis.

The Appendix is organized through two sections, where Section B.1 will review the MAP algorithm and Section B.2 will review the MLLR approach.

B.1 Maximum A Posteriori (MAP) Adaptation

The MAP approach for adjusting the parameters of a multi-parameter acoustic model in ASR was proposed in [Gauvain and Lee, 1994]. It aims to obtain a new set of the parameters in the HMM, having an initial set of them which adjust themselves quite reasonable to the correct speech model and a set of adaptation data enough to estimate new model parameters. In the end, it aims to maximize the posterior probability of the model θ to the available data x, as in Equation B.1.

$$\theta_{MAP} = \arg\max_{\theta} g(\theta|x) = \arg\max_{\theta} f(x|\theta)g(\theta) \tag{B.1}$$

After an EM procedure [Dempster et al., 1977], the final re-estimation formulas are as seen in Equations B.2, B.3, B.4, B.5 and B.6 for the means, standard deviations, weights, transition probabilities and initial probabilities respectively.

$$\hat{m}_{jk} = \frac{\tau_{jk}^{m}\mu_{jk}^{m} + \sum_{t=1}^{T}(c_{jkt}x_t)}{\tau_{jk}^{m} + \sum_{t=1}^{T}(c_{jkt})} \tag{B.2}$$

$$\hat{\sigma}_{jk} = \frac{U_{jk}^{\gamma} + S_{jk}^{\gamma} + \tau_{jk}^{m}(\mu_{jk}^{m} - \hat{m}_{jk})(\mu_{jk}^{m} - \hat{m}_{jk})^T}{\alpha_{jk}^{\gamma} - p - 1 + c_{jk}} \tag{B.3}$$

$$\hat{w}_{jk} = \frac{\gamma_{jk} - 1 + \sum_{t=1}^{T}(c_{jkt})}{\sum_{k=1}^{K}(\gamma_{jk} - 1 + \sum_{t=1}^{T} c_{jkt})} \tag{B.4}$$

$$\hat{a}_{ij} = \frac{\eta_{ij} - 1 + \sum_{t=1}^{T}(\xi_{ijt})}{\sum_{j=1}^{J}(\eta_{ij} - 1 + \sum_{t=1}^{T}(\xi_{ijt}))} \tag{B.5}$$

$$\hat{\Pi}_{i} = \frac{\eta_{i} - 1 + \gamma_{i0}}{\sum_{i=1}^{N}(\eta_{i} - 1 + \gamma_{i0})} \tag{B.6}$$

The parameters μ_{jk}^{m}, U_{jk}^{γ}, γ_{jk} and η_i are the a priori values of the mean of the Gaussian k in the state j of the current unit, the variance of the Gaussian j in the state k, of the weight of the Gaussian k in the state j and the initial weight of the unit i, respectively. On the other side, c_{jkt}, S_{jk}^{γ} and ξ_{ijt} are the probability of the frame t in Gaussian j of the state k, the averaged variance of the frames in the new adaptation data, and the number of time the Gaussian j is the most probable for the frames in the state, respectively.

The set of parameters τ_{jk}^{m} defines the weighting between the a priori data (initial model θ) and the a posteriori data (samples x) for the re-training of the Gaussian k of the state j in the unit m. The lower the value of τ, the more predominance of the new data in the new model, which can lead to a poor modeling when little data is used. On the contrary, high τ leads to a very low impact of the new data, resulting in a very little shifting of the new model with respect to the old parameters. For this work, τ was set as a constant value for all the parameters to be reestimated and was selected heuristically in the initial experiments carried out in different tasks [Saz et al., 2006b, Justo et al., 2008].

Algorithm B.1: Iterative MAP approach

```
ReadAPrioriModel($m$);
foreach Iteration $a$ do
    CalculateAPrioriParameters($\mu_{jk}^{m}$,$U_{jk}^{\gamma}$,$\gamma_{jk}$,$\eta_i$);
    foreach Input Signal $n$ do
        ForcedAlignment($n$);
        foreach Frame $t$ in $n$ assigned to unit $i$ and state $j$ do
            foreach Gaussian $k$ in $j$ do
                CalculateParameters($c_{jkt}$,$S_{jk}^{\gamma}$,$\xi_{ijt}$);
            end
        end
    end
    foreach Unit $i$ in $m$ do
        foreach State $j$ in $i$ do
            foreach Gaussian $k$ in $j$ do
                ReestimateParameters($\hat{m}_{jk}$,$\hat{\sigma}_{jk}$,$\hat{w}_{jk}$,$\hat{a}_{ij}$,$\hat{\Pi}_i$);
            end
        end
    end
end
```

The final implementation of the MAP approach in the thesis was based in an iterative framework as seen in Algorithm B.1 to achieve convergence as the models are approximating to the adaptation data. In each iteration, the input signals were aligned to the recalculated model and all the required elements for the new parameters reestimation was made. Finally, the new models parameters were obtained according to the formulas previously presented in Equations B.2 to B.6.

B.2 Maximum Likelihood Linear Regression (MLLR) Adaptation

The MLLR approach for adjusting the parameters of an acoustic model in ASR was proposed in [Legetter and Woodland, 1995]. It performs the adaptation through a set of regression matrices, which shift and rotate the a priori models towards the new parameters defined by the re-training data.

One of the main possibilities that MLLR opens is to tie a set of these matrices, so that different acoustic units can take all the data from all of them to improve the accuracy of the regression made over the input parameters. The selection of which units to tie can be made according to the similarity of the units in terms of KLD or FR of their distributions or by a priori knowledge like point or manner of articulation of the units to tie.

As argued in the main parts of the thesis, this feature of MLLR was never used in the experimental sections, because one of the targets of the thesis was to understand how phonetic mispronunciations in the speakers might produce a variation in the results in ASR. Using regression classes means that information from different units is used to retrain a given unit; this feature supposes that the knowledge of the accuracy or mistakes in the pronunciation is diluted with the information of other units.

Algorithm B.2: MLLR approach

```
if RegressionClasses then
    ReadRegressionClasses;
end
else
    DefineRegressionClasses;
end
ReadAPrioriModel($m$);
foreach Input Signal $n$ do
    ForcedAlignment($n$);
    foreach Frame $t$ in $n$ assigned to unit $i$ and state $j$ do
        StoreData;
    end
end
foreach RegressionClass $s$ do
    foreach State $j$ in $i$ do
        foreach Gaussian $k$ in $j$ do
            Calculate($G$);
            Calculate($Z$);
            CalculateRegressionMatrices($W_s$);
            PerformRegression($W_s$,$\mu_j^m$);
        end
    end
end
```

Anyways, it was evaluated if MLLR could be used without tying any units (hence, using a regression class for each single unit), but results showed up to be similar to the ones with MAP, because both are providing a similar adaptation, as they have enough data from each context to perform adaptation. For this reason, MLLR was not used extensively in the thesis and the review on MLLR is provided more briefly than in the case of MAP.

In the formulation of MLLR, a regression matrix W_s is calculated with the adaptation data through two sub-matrices G and Z as in Equation B.7 to apply the linear regression over the a priori means of the model of all Gaussians of all the states in each regression class, as in Equation B.8.

$$\bar{W_s} = G^{-1}Z \tag{B.7}$$

$$\hat{m}_j = \bar{W_s}\mu_j^m \tag{B.8}$$

For all the reasons previously mentioned, MLLR and variations of it have been used successfully in tasks for fast and reliable adaptation in case of sparsity of adaptation data.

The algorithm applied for the implementation of MLLR adaptation is depicted in Algorithm B.2.

Appendix C

Confusion Matrices for the Impaired Speakers

Life's but a walking shadow, a poor player
That struts and frets his hour upon the stage
And then is heard no more: it is a tale
Told by an idiot, full of sound and fury,
Signifying nothing.

-William Shakespeare, *The tragedy of Macbeth*

This Appendix provides the confusion matrices among the 57 words in the RFI for the 14 impaired speakers in the "Alborada-I3A" corpus. The matrices are the outcome of the ASR experiments with SI-TD models carried out in Section 4.3 and aim to show which words in the vocabulary can present a major difficulty for each speaker according to their acoustic and lexical disabilities. The Appendix is organized in the only Section C.1 which will show the matrices for all the speakers.

C.1 Confusion Matrices in the TD experiments

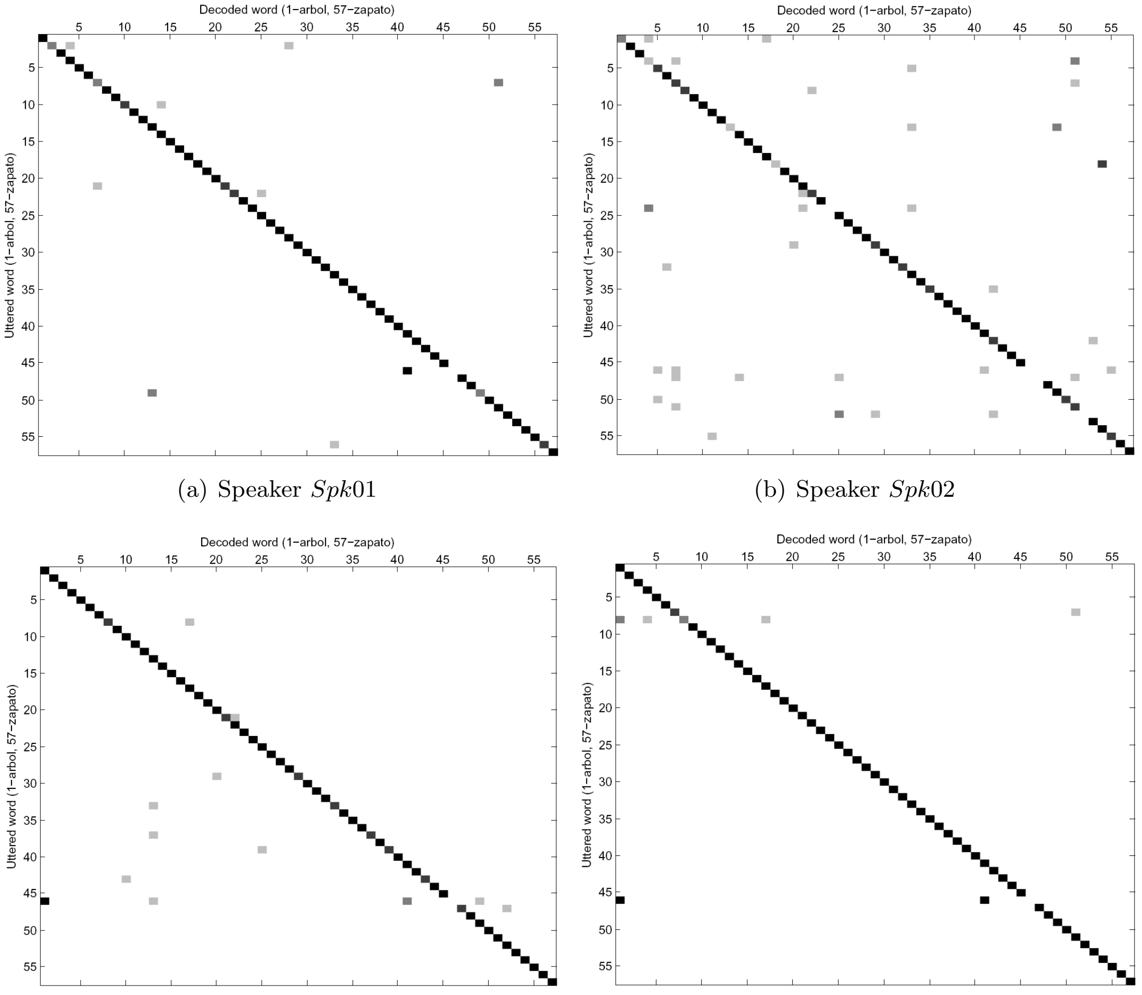
The confusion matrices for all speakers separately are shown all through Figures C.1 to C.3. These matrices were obtained from the word-unit TD ASR results in Section 4.3. and represent in the vertical axis the words as uttered by the speakers from the numbers 1 to 57 corresponding to their alphabetical order in the RFI as shown in Table 3.3 in Section 3.4. On the other side, the horizontal axis represents the word decoded by the ASR system with the same order as in the vertical axis.

Despite the lack of significance of the exact values within the confusion matrices (as there were only 4 repetitions of each word from each speaker), several trends could be studied. Speakers *Spk*01, *Spk*03, *Spk*04, *Spk*06 and *Spk*11 presented a nearly diagonal confusion matrix, corresponding to the low WER results that they achieved. On the contrary, speakers *Spk*05, *Spk*12 and *Spk*13 presented a fully blurred confusion matrix, due to their severe speech disorders and their effect on the ASR accuracy.

Although to obtain trends was difficult from the matrices, it was remarkable to see how Speakers *Spk*12 and *Spk*13 presented vertical patterns in their confusion matrices, indicating that some words were appearing very often as output of the ASR system instead of the actual uttered words:

This was the case with words 10 (*dedo*), 31 (*pala*) or 51 (*taza*) for *Spk*13 or words 8 and 10 (*clavo* and *dedo*), 17 (*globo*) or 28 and 29 (*moto* and *niño*) for *Spk*12. In nearly all cases this were 2-syllable words, following a *CVCV* structure and including at least one plosive consonant (in many cases two). All of this could be related to the possible reduction of the words as pronounced by the speakers to this phonetically simple words which would lead to the ASR errors.

These matrices showed, up to some point, the problem of ambiguity in the recognition of the speech from these users. These ambiguity would make some of the recognition mistakes irrecoverable, as the actual pronunciation of the user might be closer to the word decoded in the ASR than to the word prompted to the user (which is the ground truth of the ASR). A full phonetic labeling of the words would allow for a deeper study on the issue of ambiguity and obtain a WER over the actual pronunciation of the speakers. The cost and time demanded by a full experts' labeling limited this possibility for this thesis, but is left as a future possibility over the "Alborada-I3A" corpus.



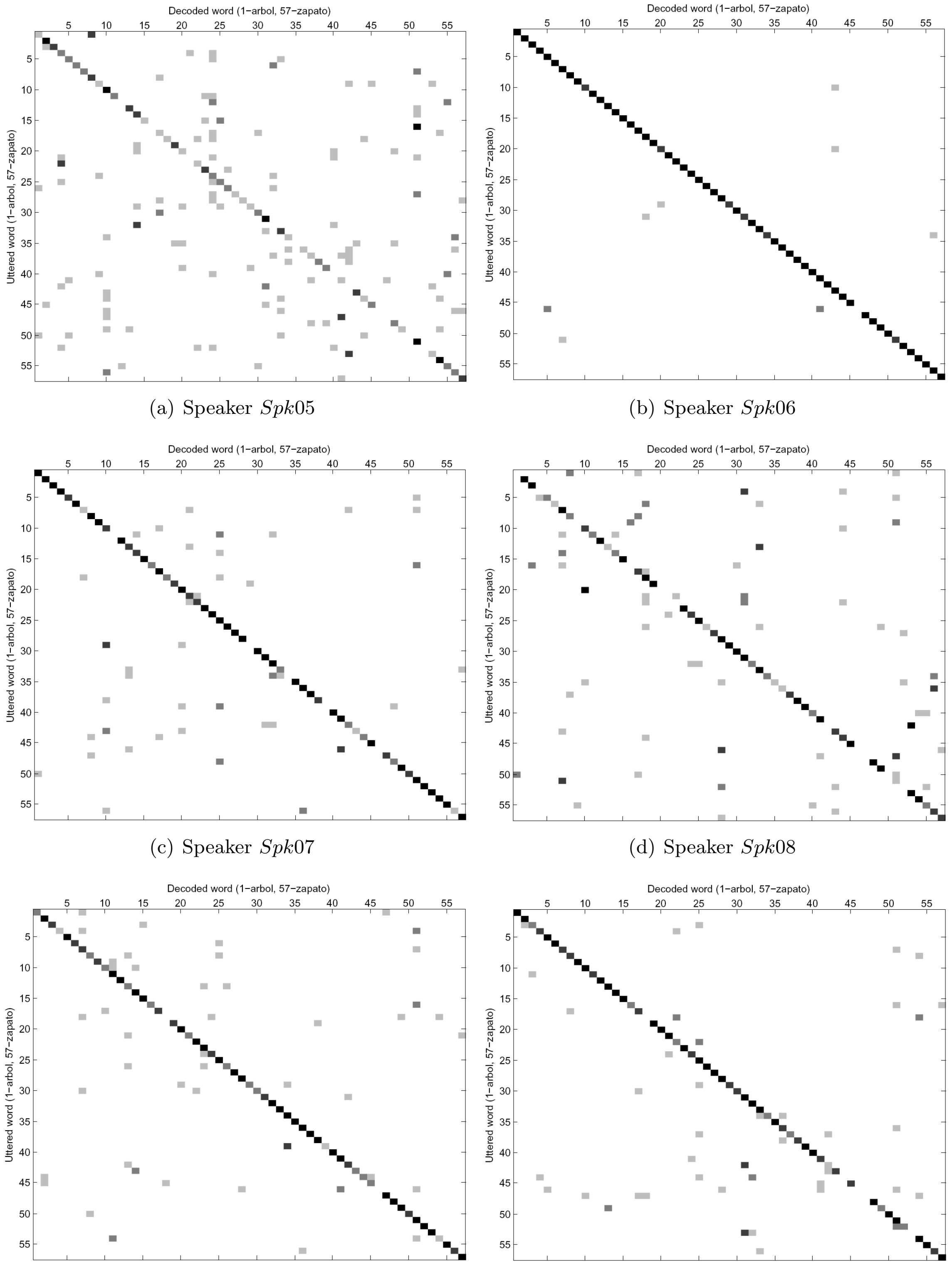
(a) Speaker *Spk*01

(b) Speaker *Spk*02

(c) Speaker *Spk*03

(d) Speaker *Spk*04

Figure C.1: Confusion matrices in the TD-ASR results: Speakers *Spk*01 to *Spk*04



(a) Speaker *Spk*05

(b) Speaker *Spk*06

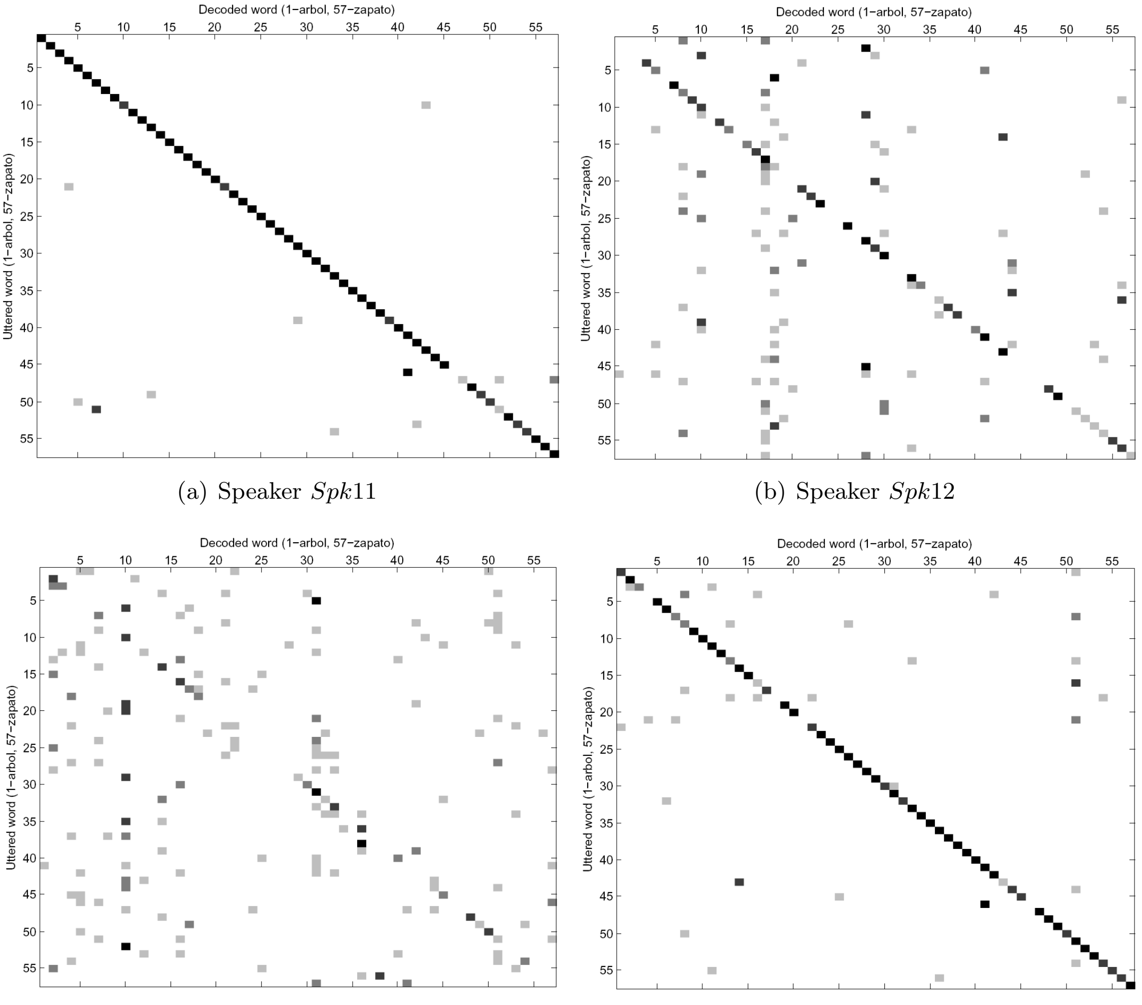
(c) Speaker *Spk*07

(d) Speaker *Spk*08

(e) Speaker *Spk*09

(f) Speaker *Spk*10

Figure C.2: Confusion matrices in the TD-ASR results: Speakers *Spk*05 to *Spk*10



(a) Speaker *Spk*11

(b) Speaker *Spk*12

(c) Speaker *Spk*13

(d) Speaker *Spk*14

Figure C.3: Confusion matrices in the TD-ASR results: Speakers *Spk*11 to *Spk*14

Bibliography

[Acero-Villán and Gomis-Cañete, 2005] Acero-Villán, P. and Gomis-Cañete, M.-J. (2005). *Tratamiento de la Voz (Manual Práctico)*. Ed. CEPE, Madrid, Spain.

[Aguado, 1999] Aguado, G. (1999). *Trastorno Específico del Lenguaje*. Ed. Aljibe, Archidona (Málaga), Spain.

[Aguinaga et al., 2004] Aguinaga, G., Armendia, M., Fraile, A., Olangua, P., and Uriz, N. (2004). *Prueba del Lenguaje Oral Navarra - revisada*. TEA S.A., Madrid, Spain.

[Alarcos, 1950] Alarcos, E. (1950). *Fonología Española*. Ed. Gredos, Madrid, Spain.

[Albor, 1991] Albor, J.-C. (1991). *ELA - Examen Logopédico de Articulación*. Ed. CEPE, Madrid, Spain.

[Alcubierre, 2005] Alcubierre, J.-M. (2005). Silla de ruedas controlada por voz: Integración, puesta a punto y pruebas de campo. Proyecto Fin de Carrera, Departamento de Ingeniería Informática y de Sistemas, University of Zaragoza, Zaragoza, Spain. Dirigido por J. Mínguez y L. Montesano (Ponente L. Montano).

[Alcubierre et al., 2005] Alcubierre, J.-M., Mínguez, J., Montesano, L., Montano, L., Saz, O., and Lleida, E. (2005). Silla de ruedas inteligente controlada por voz. In *Proceedings of the Primer Congreso Internacional de Domótica, Robótica y Teleasistencia para todos*, pages 349–360, Madrid, Spain.

[Amdal et al., 2009] Amdal, I., Johnsen, M.-H., and Versvik, E. (2009). Automatic evaluation of quantity contrast in non-native Norwegian speech. In *Proceedings of the 2009 Workshop on Speech and Language Technologies in Education (SLaTE)*, Wroxall Abbey Estates, United Kingdom.

[Atwell et al., 2003] Atwell, E., Howarth, P., and Souter, C. (2003). The ISLE corpus: Italian and German spoken learners' English. *ICAME JOURNAL - Computers in English Linguistics*, 27:5–18.

[Auckenthaler et al., 2000] Auckenthaler, R., Carey, M., and Lloyd-Thomas, H. (2000). Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1–3):42–54.

[Bergua, 2005] Bergua, B. (2005). Desarrollo de un ratón virtual configurable sobre el S.O. Windows. Proyecto Fin de Carrera, Departamento de Ingeniería Informática y de Sistemas, University of Zaragoza, Zaragoza, Spain. Dirigido por J. Ezpeleta.

[Bergua et al., 2006] Bergua, B., Marcos, J.-M., and Canalís, C. (2006). Ratón virtual: Ayuda técnica desarrollada a partir del acuerdo de colaboración entre el C.E.E. "Alborada" y el Centro Politécnico Superior de la Universidad de Zaragoza. In *Proceedings of the Jornadas Nacionales de Sistemas Aumentativos de Comunicación*, Zaragoza, Spain.

[Berkson, 2005] Berkson, G. (2005). Intellectual and physical disabilities in prehistory and early civilization. *Mental Retardation*, 42(3):195–208.

[Beukelman and Mirenda, 1998] Beukelman, D.-R. and Mirenda, P. (1998). *Augmentative and Alternative Communication: Management of Severe Communication Disorders in Children and Adults.* Brookes Publishing Company, Baltimore (MD), USA.

[Bilmes et al., 2006] Bilmes, J., Malkin, J., Li, X., Harada, S., Kilanski, K., Kirchhoff, K., Wright, R., Subramanya, A., Landay, J., Dowden, P., and Chizeck, H. (2006). The vocal joystick. In *Proceedings of the 2006 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 625–628, Toulouse, France.

[Black et al., 2008] Black, M., Tepperman, J., Kazemzadeh, A., Lee, S., and Narayanan, S. (2008). Pronunciation verification of english letter-sounds in preliterate children. In *Proceedings of the 10th International Conference on Spoken Language Processing (ICSLP - Interspeech)*, pages 2783–2786, Brisbane, Australia.

[Blaney and Wilson, 2000] Blaney, B. and Wilson, J. (2000). Acoustic variability in dysarthria and computer speech recognition. *Clinical Linguistics & Phonetics*, 14(4):307–327.

[Bocklet et al., 2009] Bocklet, T., Haderlein, T., Honig, F., Rosanowski, F., and Noth, E. (2009). Evaluation and assessment of speech intelligibility on pathologic voices based upon acoustic speaker models. In *Proceedings of 3rd Advanced Voice Function Assessment International Workshop (AVFA09)*, pages 89–92, Madrid, Spain.

[Bosch-Galcerán, 2004] Bosch-Galcerán, L. (2004). *Evaluación Fonólogica del Habla Infantil.* Ed. Masson, Barcelona, Spain.

[Buera et al., 2007] Buera, L., Lleida, E., Miguel, A., Ortega, A., and Saz, O. (2007). Cepstral vector normalization based on stereo data for robust speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1098–1113.

[Bustos, 1995] Bustos, M.-C. (1995). *Manual de Logopedia Escolar. Niños con Alteraciones del Lenguaje Oral en Educación Infantil y Primaria.* Ed. CEPE, Madrid, Spain.

[Caballero et al., 2002] Caballero, M., Mariño, J., and Moreno, A. (2002). Multidialectal Spanish modeling for ASR. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, Las Palmas de Gran Canaria, Spain.

[Caballero et al., 2004] Caballero, M., Moreno, A., and Nogueiras, A. (2004). Data driven multidialectal phone set for Spanish dialects. In *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP - Interspeech)*, pages 837–840, Jeju Island, Korea.

[Calero-Vaquera and Calvillo-Jurado, 1992] Calero-Vaquera, M.-L. and Calvillo-Jurado, M. (1992). Consideraciones sobre el yeísmo en la enseñanza del español. *Cauce*, 14–15:37–46.

[Caplan, 1987] Caplan, D. (1987). *Neurolinguistics and Linguistic Aphasiology: An Introduction.* Cambridge University Press, Cambridge, UK.

[Chen et al., 2000] Chen, K.-T., Liau, W.-W., Wang, H.-M., and Lee, L.-S. (2000). Fast speaker adaptation using eigenspace-based Maximum Likelihood Linear Regression. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP - Interspeech)*, pages 742–745, Beijing, China.

[Chou, 2005] Chou, F.-C. (2005). Ya-Ya language box - A portable device for English pronunciation training with speech recognition technologies. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 169–172, Lisbon, Portugal.

[Cleuren et al., 2006] Cleuren, L., Duchateau, J., Sips, A., Ghesquiere, P., and Hamme, H. V. (2006). Developing an automatic assessment tool for childrens oral reading. In *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP - Interspeech)*, pages 817–820, Pittsburgh (PA), USA.

[Conrad, 1907] Conrad, J. (1907). *The Secret Agent: A Simple Tale.* Methuen & Co, London, United Kingdom.

[Cover and Thomas, 1991] Cover, T.-M. and Thomas, J.-A. (1991). *Elements on Information Theory.* Wiley Interscience, New York (NY), USA.

[Creer et al., 2009] Creer, S., Cunningham, S.-P., Green, P.-D., and Fatema, K. (2009). Personalizing synthetic voices for people with progressive speech disorders: Judging voice similarity. In *Proceedings of the 11th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 1427–1430, Brighton, UK.

[Creer et al., 2010] Creer, S., Green, P., Cunningham, S., and Yamaghisi, J. (2010). Building personalized synthesized voices for individuals with dysarthria using the HTK toolkit. In Mullenix, J.-W. and Stern, S.-E., editors, *Computer Synthesized Speech Technologies: Tools for Aiding Impairment.* IGI Publishing, Hershey (PA), USA.

[Croot, 1999] Croot, K. (1999). An acoustic analysis of vowel production across tasks in a case of non-fluent progressive aphasia. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP-Interspeech)*, pages 907–910, Sydney, Australia.

[Croot et al., 2000] Croot, K., Hodges, J.-R., Xuereb, J., and Patterson, K. (2000). Phonological and articulatory impairment in alzheimer's disease: A case series. *Brain and Language*, 75:277–309.

[Cucchiarini et al., 2008a] Cucchiarini, C., Lembrechts, D., and Strik, H. (2008a). Hlt and communicative disabilities: The need for co-operation between government, industry and academia. In *Proceedings of LangTech 2008*, pages 125–128, Rome, Italy.

[Cucchiarini et al., 2007] Cucchiarini, C., Neri, A., de Wet, F., and Strik, H. (2007). ASR-based pronunciation training: Scoring accuracy and pedagogical effectiveness of a system for Dutch L2 learners. In *Proceedings of the 10th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 2181–2184, Antwerp, Belgium.

[Cucchiarini et al., 2008b] Cucchiarini, C., van Doremalen, J., and Strik, H. (2008b). DISCO: Development and Integration of Speech technology into COurseware for language learning. In *Proceedings of the 2008 International Conference on Spoken Language Processing (ICSLP - Interspeech)*, pages 2791–2794, Brisbane, Australia.

[Cylwik et al., 2009] Cylwik, N., Wagner, A., and Demenko, G. (2009). The EURONOUNCE corpus of non-native Polish for ASR-based pronunciation tutoring system. In *Proceedings of the 2009 Workshop on Speech and Language Technologies in Education (SLaTE)*, Wroxall Abbey Estates, United Kingdom.

[Deller et al., 1991] Deller, J.-R., Hsu, D., and Ferrier, L.-J. (1991). On the use of Hidden Markov Modelling for recognition of dysarthric speech. *Computer Methods and Programs in Biomedicine*, 35:125–139.

[Deller et al., 1993] Deller, J.-R., Liu, M.-S., Ferrier, L.-J., and Robichaud, P. (1993). The whitaker database of dysarthric (cerebral palsy) speech. *Journal of the Acoustical Society of America*, 93(6):3516–3518.

[Dempster et al., 1977] Dempster, A.-P., Laird, N.-M., and Rubin, D.-B. (1977). Maximum Likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–21.

[D'Haro et al., 2008] D'Haro, L.-F., San-Segundo, R., de Córdoba, R., Bungeroth, J., Stein, D., and Ney, H. (2008). Language model adaptation for a speech to sign language translation system using web frequencies and a map framework. In *Proceedings of the 10th International Conference on Spoken Language Processing (ICSLP-Interspeech)*, pages 2199–2202, Brisbane, Australia.

[Disordered Voice Database, 1994] Disordered Voice Database (1994). Disordered voice database v.1.03. Boston (MA), USA.

[Duchateau et al., 2007] Duchateau, J., Cleuren, L., Hamme, H. V., and Ghesquiere, P. (2007). Automatic assessment of children's reading level. In *Proceedings of the 10th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 1210–1213, Antwerp, Belgium.

[Ephrain and Malah, 1985] Ephrain, Y. and Malah, D. (1985). Speech enhancement using a Minimum Mean-Square Error Log-Spectral Amplitude estimator. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 33(2):443–445.

[Escartín, 2008] Escartín, A. (2008). Gestión de "COMUNICA: Conjunto de herramientas para la logopedia" y ampliación de sus herramientas a los niveles semántico y pragmático del lenguaje. Proyecto Fin de Carrera, Departamento de Ingeniería Electrónica y Comunicaciones, University of Zaragoza, Zaragoza, Spain. Dirigido por O. Saz (Ponente E. Lleida).

[Falcó et al., 2006] Falcó, J., Plaza, I., Marcos, J.-M., and Canalís, C. (2006). Dispositivo de orientación temporal: Ayuda técnica desarrollada a partir del acuerdo de colaboración entre el C.E.E. "Alborada" y el Centro Politécnico Superior de la Universidad de Zaragoza. In *Proceedings of the Jornadas Nacionales de Sistemas Aumentativos de Comunicación*, Zaragoza, Spain.

[Fant, 1960] Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton & Co., Den Haague, The Netherlands.

[Fengpei et al., 2008] Fengpei, G., Fuping, P., Changliang, L., Bin, D., and Yonghong, Y. (2008). Forward optimal modeling of acoustic confusions in mandarin call system. In *Proceedings of the 2008 International Conference on Spoken Language Processing (ICSLP - Interspeech)*, pages 2807–2810, Brisbane, Australia.

[Ferrier et al., 1995] Ferrier, L.-J., Shane, H.-C., Ballard, H.-F., Carpenter, T., and Benoit, A. (1995). Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative and Alternative Communication*, 11(3):165–175.

[Flora, 1987] Flora, P. (1987). *Growth to Limits: The Western European Welfare States Since World War II*. Walter De Gruyter Inc, Berlin, Germany.

[Fossler-Lussier et al., 2005] Fossler-Lussier, E., Rytter, C.-A., and Srinivasan, S. (2005). Phonetic ignorance is bliss: Investigating the effects of phonetic information reduction on ASR performance. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 1249–1252, Lisbon, Portugal.

[Frago-Gracia, 1978] Frago-Gracia, J.-A. (1978). La actual irrupción del yeísmo en el español navarroaragonés y otras cuestiones históricas. *Archivo de Filología Aragonesa*, 22–23:7–19.

[Fry, 1958] Fry, D. (1958). Experiments in perception of stress. *Language and Speech*, 1:126–152.

[García et al., 2008] García, J.-E., Ortega, A., Miguel, A., and Lleida, E. (2008). Sistema de reconocimiento automático del habla distribuido aplicado a entornos logísticos. In *Proceedings of the V Jornadas en Tecnologías del Habla*, pages 240–243, Bilbao, Spain.

[García-Gómez et al., 1999] García-Gómez, R., López-Barquilla, R., Puertas-Tera, J.-I., Parera-Bermúdez, J., Haton, M.-C., Haton, J.-P., Alinat, P., Moreno, S., Hess, W., Sánchez-Raya, M.-A., Martínez-Gual, E.-A., Navas-Chabeli-Daza, J. L., Antoine, C., Durel, M.-M., Maurin, G., and Hohmann, S. (1999). Speech training for deaf and hearing impaired people: ISAEUS consortium. In *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 1067–1070, Budapest, Hungary.

[Garland, 1995] Garland, R. (1995). *The Eye of the Beholder: Deformity and Disability in the Graeco-Roman World.* Cornell University Press, Ithaca (NY), USA.

[Garofalo et al., 1993] Garofalo, J.-S., Lamel, L.-F., Fisher, W.-M., Fiscus, J.-G., Pallett, D.-S., Dahlgren, N.-L., and Zue, V. (1993). TIMIT acoustic-phonetic continuous speech corpus. Technical report, Linguistic Data Consortium, Philadelphia (PA), USA.

[Gauvain and Lee, 1994] Gauvain, J.-L. and Lee, C.-H. (1994). Maximum A Posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298.

[Gerosa and Narayanan, 2008] Gerosa, M. and Narayanan, S. (2008). Investigating assessment of reading comprehension in young children. In *Proceedings of the 2008 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5057–5060, Las Vegas (NV), USA.

[Godino-Llorente and Gómez-Vilda, 2004] Godino-Llorente, J.-I. and Gómez-Vilda, P. (2004). Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *IEEE Transactions on Biomedical Engineering*, 51(2):380–384.

[Gómez et al., 2005] Gómez, P., Lázaro, C., Fernández, R., Nieto, A., Godino, J.-I., Martínez, R., Díaz, F., Álvarez, A., Murphy, K., Nieto, V., Rodellar, V., and Fernández, F.-J. (2005). Using biomechanical parameter estimates in voice pathology detection. In *Proceedings of 4th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MABEVA05)*, pages 29–31, Florence, Italy.

[González, 2009] González, I. (2009). MICE: Entorno para la integración de componentes para el control de ratón. Proyecto Fin de Carrera, Departamento de Ingeniería Informática y de Sistemas, University of Zaragoza, Zaragoza, Spain. Dirigido por J. Ezpeleta.

[Gracia et al., 2009] Gracia, A., Arsuaga, J.-L., Martínez, I., Lorenzo, C., Carretero, J.-M., de Castro, J.-M. B., and Carbonell, E. (2009). Craniosynostosis in the middle pleistocene human cranium 14 from the sima de los huesos, atapuerca, spain. *Proceedings of the National Academy of Sciences (PNAS)*, 106:6573–6578.

[Granstroem, 2005] Granstroem, B. (2005). Speech technology for language training and e-inclusion. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 449–452, Lisbon, Portugal.

[Green et al., 2003] Green, P., Carmichael, J., Hatzis, A., Enderby, P., Hawley, M., and Parker, M. (2003). Automatic Speech Recognition with sparse training data for dysarthric speakers. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 1189–1192, Geneva, Switzerland.

[Hansen, 1999] Hansen, J.-H.-L. (1999). SUSAS. Technical report, Linguistic Data Consortium, Philadelphia (PA), USA.

[Harada et al., 2008] Harada, S., Landay, J., Malkin, J., Li, X., and Bilmes, J. (2008). The Vocal Joystick: evaluation of voice-based cursor control techniques for assistive technology. *Disability and Rehabilitation: Assistive Technology*, 3(1):22–34.

[Hatzis, 1999] Hatzis, A. (1999). *Optical Logo-Therapy: Computer-Based Audio-Visual Feedback Using Interactive Visual Displays for Speech Training.* PhD thesis, Department of Computer Science, The University of Sheffield, Sheffield, United Kingdom.

[Hatzis et al., 2003] Hatzis, A., Green, P., Carmichael, J., Cunningham, S., Palmer, R., Parker, M., and O'Neill, P. (2003). An integrated toolkit deploying speech technology for computer based speech training with application to dysarthric speakers. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 2213–2216, Geneva, Switzerland.

[Hatzis et al., 1999] Hatzis, A., Green, P.-D., and Howard, S. (1999). Optical-Logo-Therapy (OLT): Visual displays in practical auditory phonetics teaching. Technical report, Department of Computer Science, Department of Human Communication Science; University of Sheffield.

[Hatzis et al., 1997] Hatzis, A., Green, P.-D., and Howard, S.-J. (1997). Optical Logo-Therapy (OLT) : A computer-based real time visual feedback application for speech training. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 1763–1766, Rhodes, Greece.

[Hawley et al., 2007] Hawley, M., Cunningham, S., Cardinaux, F., Coy, A., O'Neill, P., Seghal, S., and Enderby, P. (2007). Challenges in developing a voice input voice output communication aid for people with severe dysarthria. In Eizmendi, G., Azkoitia, J.-M., and Craddock, G.-M., editors, *Challenges for Assistive Technology*, pages 363–367. IOS Press, Amsterdam, The Netherlands.

[Hawley et al., 2003] Hawley, M., Enderby, P., Green, P., Brownsell, S., Hatzis, A., Parker, M., Carmichael, J., Cunningham, S., O'Neill, P., and Palmer, R. (2003). STARDUST Speech Training And Recognition for Dysarthric Users of aSsistive Technology. In *Proceedings of the 7th Conference of the Association for the Advancement of Assistive Technology in Europe, AAATE*, Dublin, Ireland.

[Hawley et al., 2005] Hawley, M.-S., Green, P., Enderby, P., Cunningham, S., and Moore, R.-K. (2005). Speech technology for e-inclusion of people with physical disabilities and disordered speech. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 445–448, Lisbon, Portugal.

[Hirano, 1981] Hirano, M. (1981). Psycho-acoustic evaluation of voice: GRBAS scale for evaluating the hoarse voice. In Hirano, M., editor, *Clinical Examination of voice*. Springer Verlag, New York City (NY), USA.

[Hirsch and Pearce, 2000] Hirsch, H.-G. and Pearce, D. (2000). The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proceedings of the ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenged for the Next Millenium"*, pages 18–20, Paris, France.

[Hosom et al., 2003] Hosom, J.-P., Kain, A., Mishra, T., J.-P.-H. van Santen, M. F.-O., and Staehely, J. (2003). Intelligibility of modifications to dysarthric speech. In *Proceedings of the 2005 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 924–927, Hong Kong, China.

[Hualde, 2005] Hualde, J.-I. (2005). *The Sounds of Spanish.* Cambridge University Press, Cambridge, United Kingdom.

[Huang et al., 2001] Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing.* Prentice Hall, Upper Saddle River (NJ), USA.

[Iturriate et al., 2009] Iturriate, I., Antelis, and Mínguez, J. (2009). Synchronous EEG brain-actuated wheelchair with automated navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Kobe, Japan.

[Jacobson et al., 1997] Jacobson, B., Johnson, A., Grywalski, C., and Silbergleit, A. (1997). The Voice Handicap Index (VHI): development and validation. *American Journal of Speech and Language Pathology*, 6(1):66–69.

[Janin et al., 2004] Janin, A., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2004). ICSI meeting speech. Technical report, Linguistic Data Consortium, Philadelphia (PA), USA.

[Jelinek, 1998] Jelinek, F. (1998). *Statistical Methods For Speech Recognition.* MIT Press, Cambridge (MA), USA.

[Justo et al., 2008] Justo, R., Saz, O., Guijarrubia, V., Miguel, A., Torres, M.-I., and Lleida, E. (2008). Improving dialogue systems in a home automation environment. In *Proceedings of the First International Conference on Ambient Media and Systems (Ambi-Sys 2008)*, Québec City, Canada.

[Kain et al., 2004] Kain, A., Niu, X., Hosom, J.-P., Miao, Q., and van Santen, J. (2004). Formant re-synthesis of dysarthric speech. In *Proceedings of the 5th ISCA Speech Synthesis Workshop*, pages 25–30, Pittsburgh (PA), USA.

[Kanters et al., 2009] Kanters, S., Cucchiarini, C., and Strik, H. (2009). The Goodness of Pronunciation algorithm: a detailed performance study. In *Proceedings of the 2009 Workshop on Speech and Language Technologies in Education (SLaTE)*, Wroxall Abbey Estates, United Kingdom.

[Kenny et al., 2006] Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P. (2006). Improvements in factor analysis-based speaker verification. In *Proceedings of the 2006 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 113–116, Toulouse, France.

[Kenny et al., 2003] Kenny, P., Mihoubi, M., and Dumouchel, P. (2003). New MAP estimators for speaker recognition. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 2961–2964, Geneva, Switzerland.

[Kim et al., 2008] Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., and France, S. (2008). Dysarthric speech database for universal access research. In *Proceedings of the 10th International Conference on Spoken Language Processing (ICSLP-Interspeech)*, pages 1741–1744, Brisbane, Australia.

[Koehn, 2005] Koehn, P. (2005). Europarl: A parallel corpus for statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, Phuket,Thailand.

[Koul, 2003] Koul, R.-K. (2003). Synthetic speech perception in individuals with and without disabilities. *Augmentative and Alternative Communication*, 19:49–58.

[Koul and Clapsaddle, 2003] Koul, R.-K. and Clapsaddle, K.-C. (2003). Effects of repeated listening experiences on the perception of synthetic speech by individuals with mild-to-moderate intellectual disabilities. *Augmentative and Alternative Communication*, 22:1–11.

[Koul and Hanners, 1997] Koul, R.-K. and Hanners, J. (1997). Word identification and sentence verification of two synthetic speech systems by individuals with intellectual disabilities. *Augmentative and Alternative Communication*, 13:99–107.

[Kozma, 2005] Kozma, C. (2005). Dwarfs in ancient egypt. *American Journal of Medical Genetics Part A*, 140A(4):303–311.

[Launa and Borel-Maisonny, 1989] Launa, Y.-C. and Borel-Maisonny, S. (1989). *Trastornos del Lenguaje, la Palabra y la Voz en el Niño*. Ed. Masson, Barcelona, Spain.

[Lee and Seneff, 2006] Lee, J. and Seneff, S. (2006). Automatic grammar correction for second-language learners. In *Proceedings of the 2006 International Conference on Spoken Language Processing (ICSLP - Interspeech)*, pages 1978–1981, Pittsburgh (PA), USA.

[Lee, 1989] Lee, K.-F. (1989). *Automatic Speech Recognition: The Development of the Sphinx System*. Kluwer Academic Publishers, Norwell (MA), USA.

[Lee and Hon, 1989] Lee, K.-F. and Hon, H.-W. (1989). Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(11):1641–1648.

[Lefévre, 1996] Lefévre, J.-P. (1996). HARP: An autonomous rehabilitation system for hearing impaired people. Technical report, TIDE Project 1060.

[Legetter and Woodland, 1995] Legetter, C.-J. and Woodland, P.-C. (1995). Maximum Likelihood Linear Regression for speaker adaptation of the parameters of continous density Hidden Markov Models. *Computer Speech and Language*, 9:171–185.

[Leonard and Doddington, 1993] Leonard, R.-G. and Doddington, G. (1993). TIDIGITS. Technical report, Linguistic Data Consortium, Philadelphia (PA), USA.

[Liu, 1996] Liu, S.-A. (1996). Landmark detection for distinctive feature-based speech recognition. *Journal of the Acoustic Society of America*, 100(5):3417–3430.

[Lleida and Rose, 2000] Lleida, E. and Rose, R.-C. (2000). Utterance verification in continuous speech recognition: Decoding and training procedures. *IEEE Transactions on Speech and Audio Processing*, 8(2):126–139.

[Luo et al., 2008] Luo, D., Shimomura, N., Minematsu, N., Yamauchi, Y., and Hirose, K. (2008). Automatic pronunciation evaluation of language learners utterances generated through shadowing. In *Proceedings of the 2008 International Conference on Spoken Language Processing (ICSLP - Interspeech)*, pages 2807–2810, Brisbane, Australia.

[Mak et al., 2003] Mak, B., Siu, M., Ng, M., Tam, Y.-C., Y.-C., Chany, Chan, K.-W., Leung, K.-Y., Ho, S., Chong, F.-H., Wong, J., and Lo, J. (2003). PLASER: Pronunciation learning via Automatic Speech Recognition. In *Proceedings of the HLT-NAACL 03 Workshop on Building educational applications using natural language processing*, pages 23–29, Edmonton (AL), Canada.

[Martin and Przybocki, 2004] Martin, A. and Przybocki, M. (2004). 2002 NIST Speaker Recognition Evaluation. Technical report, Linguistic Data Consortium, Philadelphia (PA), USA.

[Martínez et al., 2007] Martínez, B., Peguero, P., Ezpeleta, J., Falcó, J., Lleida, E., Mínguez, J., and Saz, O. (2007). Universidad y educación especial: Desarrollo y resultados de la colaboración entre el Centro Politécnico Superior y el Centro de Educación Especial "Alborada". In *Proceedings of the III Congreso Nacional sobre Universidad y Discapacidad*, Zaragoza, Spain.

[Martínez-Celdrán and Fernández-Planas, 2007] Martínez-Celdrán, E. and Fernández-Planas, A.-M. (2007). *Manual de Foética Espñola: Articulaciones y Sonidos del Espñol*. Ed. Ariel, Madrid, Spain.

[Marujo et al., 2009] Marujo, L., Lopes, J., Mamede, N., Trancoso, I., Pino, J., Eskenazi, M., baptista, J., and Viana, C. (2009). Semi-automatic generation of cloze question distractors effect of students' L1. In *Proceedings of the 2009 Workshop on Speech and Language Technologies in Education (SLaTE)*, Wroxall Abbey Estates, United Kingdom.

[McGraw et al., 2009] McGraw, I., Yoshimoto, B., and Seneff, S. (2009). Speech-enabled card games for incidental vocabulary acquisition in a foreign language. *Speech Communication*, 51(10):1006–1023.

[Menéndez-Pidal et al., 1996] Menéndez-Pidal, X., Polikoff, J.-B., Peters, S.-M., Lorenzo, J., and Bunnell, H.-T. (1996). The Nemours database of dysarthric speech. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP-Interspeech)*, pages 1962–1965, Philadelphia (PA), USA.

[Miguel et al., 2008] Miguel, A., Lleida, E., Rose, R., Buera, L., Saz, O., and Ortega, A. (2008). A normalization model for capturing local variability in speaker independent ASR. *IEEE Transactions on Audio, Speech and Language Processing*, 16(3):578–593.

[Mínguez and Montano, 2005] Mínguez, J. and Montano, L. (2005). Sensor-based robot motion generation in unknown, dynamic and troublesome scenarios. *Robotics and Autonomous Systems*, 52(4):290–311.

[Mínguez et al., 2006] Mínguez, J., Montesano, L., and Montano, L. (2006). Autonomous motion generation for a robotic wheelchair. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Orlando (FLA), USA.

[Monfort and Juárez-Sánchez, 1989] Monfort, M. and Juárez-Sánchez, A. (1989). *Registro Fonológico Inducido (Tarjetas Gráficas)*. Ed. Cepe, Madrid, Spain.

[Monfort and Monfort-Juárez, 2001a] Monfort, M. and Monfort-Juárez, I. (2001a). *En la Mente 2. ¿Cómo decirlo? Un Soporte Gráfico para el Entrenamiento de las Habilidades Pragmáticas en el Niño*. Entha ediciones, Madrid, Spain.

[Monfort and Monfort-Juárez, 2001b] Monfort, M. and Monfort-Juárez, I. (2001b). *En la Mente. Un Soporte Gráfico para el Entrenamiento de las Habilidades Pragmáticas en el Niño*. Entha ediciones, Madrid, Spain.

[Moore and ten Bosch, 2009] Moore, R. and ten Bosch, L. (2009). Modeling vocabulary growth from birth to young adulthood. In *Proceedings of the 11th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 1727–1730, Brighton, UK.

[Moreno et al., 2000] Moreno, A., Lindberg, B., Draxler, C., Richard, G., Choukri, K., Euler, S., and Allen, J. (2000). Speech Dat Car. A large speech database for automotive environments. In *Proceedings of the II Language Resources European Conference*, Athens, Greece.

[Moreno et al., 1993] Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J.-B., and Nadeu, C. (1993). Albayzin speech database: Design of the phonetic corpus. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 175–178, Berlin, Germany.

[Morley, 1994] Morley, J. (1994). *Pronunciation pedagogy and theory: New view, new directions.* TESOL Publications, Alexandria (VA), USA.

[Navarro-Mesa et al., 2005] Navarro-Mesa, J.-L., Quintana-Morales, P., Pérez-Castellano, I., and Espinosa-Yáñez, J. (2005). Oral corpus of the project HACRO (help tool for the confidence of oral utterances). Technical report, Department of Signal and Communications, University of Las Palmas de Gran Canaria.

[Negre, 2005] Negre, F. (2005). Desarrollo de herramientas para la creación y utilización de tableros de comunicación en el ámbito de la educación especial [recurso electrónico]. Proyecto Fin de Carrera, Departamento de Ingeniería Informática y de Sistemas, , University of Zaragoza, Zaragoza, Spain. Dirigido por J. Ezpeleta.

[Negre et al., 2006] Negre, F., Ramos, D., Marcos, J.-M., and Canalís, C. (2006). Generador interactivo de tableros de comunicación: Ayuda técnica desarrollada a partir del acuerdo de colaboración entre el C.E.E. "Alborada" y el Centro Politécnico Superior de la Universidad de Zaragoza. In *Proceedings of the Jornadas Nacionales de Sistemas Aumentativos de Comunicación*, Zaragoza, Spain.

[Neiberg et al., 2008] Neiberg, D., Ananthakrishnan, G., and Engwall, O. (2008). The acoustic to articulation mapping: Non-linear or non-unique? In *Proceedings of the 10th International Conference on Spoken Language Processing (ICSLP-Interspeech)*, pages 1485–1488, Brisbane, Australia.

[Neri et al., 2006] Neri, A., Cucchiarini, C., and Strik, H. (2006). Improving segmental quality in L2 Dutch by means of Computer Assisted Pronunciation Training with Automatic Speech Recognition. In *Proceedings of CALL 2006*, pages 144–151, Antwerp, Belgium.

[Neugebauer, 1989] Neugebauer, R. (1989). Diagnosis, guardianship, and residential care of the mentally ill in medieval and early modern england. *American Journal of Psychiatry*, 146(12):1580–1584.

[Oester et al., 2003] Oester, A.-M., House, D., Hatzis, A., and Green, P. (2003). Testing a new method for training fricatives using visual maps in the Ortho-Logo-Paedia project (OLP). In *Proceedings of the XVI Swedish Phonetics Conference (Fonetik 2003)*, Umea, Sweden.

[Oester et al., 2002] Oester, A.-M., House, D., Protopapas, A., and Hatzis, A. (2002). Presentation of a new EU project for speech therapy: OLP (Ortho-Logo-Paedia). In *Proceedings of the XV Swedish Phonetics Conference (Fonetik 2002)*, pages 45–48, Stockholm, Sweden.

[Ortega et al., 2009] Ortega, A., García, J.-E., Miguel, A., and Lleida, E. (2009). Real-time live broadcast news subtitling system for spanish. In *Proceedings of the 11th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, Brighton, United Kingdom.

[Ortega et al., 2004] Ortega, A., Sukno, F., Lleida, E., Miguel, A., and Buera, L. (2004). AV@CAR: A Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition. In *Proceedings of the 4th International Conferencece on Language Resources and Evaluation*, pages 763–767, Lisbon, Portugal.

[Pantazis et al., 2009] Pantazis, Y., Rosec, O., and Stylianou, Y. (2009). Time-frequency representation of pathologic signals based on a quasi harmonic model of speech. In *Proceedings of 3rd Advanced Voice Function Assessment International Workshop (AVFA09)*, pages 157–160, Madrid, Spain.

[Pascual-García, 1992] Pascual-García, P. (1992). *La dislalia: Naturaleza, Diagnóstico, Rehabilitación.* Ed. Cepe, Madrid, Spain.

[Patel, 2002] Patel, R. (2002). Phonatory control in adults with cerebral palsy and severe dysarthria. *Augmentative and Alternative Communication*, 18:2–10.

[Perelló, 1984] Perelló, P. (1984). *Trastornos del Habla y Trastornos del Lenguaje.* Ed. Científico-Médica, Barcelona, Spain.

[Pino and Eskenazi, 2009] Pino, J. and Eskenazi, M. (2009). Semi-automatic generation of cloze question distractors effect of students' l1. In *Proceedings of the 2009 Workshop on Speech and Language Technologies in Education (SLaTE)*, Wroxall Abbey Estates, United Kingdom.

[Potamianos and Neti, 2001] Potamianos, G. and Neti, C. (2001). Automatic speechreading of impaired speech. In *Proceedings of the International Conference on Auditory-Visual Speech Processing*, pages 177–182, Aalborg, Denmark.

[Prizl-Jakovac, 1999] Prizl-Jakovac, T. (1999). Vowel production in aphasia. In *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 583–586, Budapest, Hungary.

[Quilis, 1981] Quilis, A. (1981). *Fonética Acústica de la Lengua Española.* Ed. Gredos, Madrid, Spain.

[Rapin and Allen, 1983] Rapin, I. and Allen, D. (1983). Developmental language disorders: Nosologic considerations. In Kirk, U., editor, *Neuropsychology of Language, Reading, and Spelling*, pages 155–184. Academic Press, New York City (NY), USA.

[Rodríguez, 2008] Rodríguez, V. (2008). El uso de herramientas multimedia para la práctica de la pronunciación en clases de ele con adolescentes. Memoria final del Máster en Enseñanza del Español como Lengua Extranjera (MEELE). Universidad Antonio de Nebrija, Departamento de Lenguas Aplicadas.

[Rodríguez and Lleida, 2009] Rodríguez, W.-R. and Lleida, E. (2009). Formant estimation in children's speech and its application for a spanish speech therapy tool. In *Proceedings of the 2009 Workshop on Speech and Language Technologies in Education (SLaTE)*, Wroxall Abbey Estates, United Kingdom.

[Rodríguez et al., 2008a] Rodríguez, W.-R., Saz, O., Lleida, E., Vaquero, C., and Escartín, A. (2008a). Comunica - tools for speech and language therapy. In *Proceedings of the 2008 Workshop on Children, Computer and Interaction*, Chania, Greece.

[Rodríguez et al., 2009] Rodríguez, W.-R., Saz, O., Vaquero, C., and Lleida, E. (2009). Habilitación del prelenguaje y del lenguaje con Comunica. In *Proceedings of the VIII Congreso Iberoamericano de Informática y Educación Especial (CIIEE)*, San José, Costa Rica.

[Rodríguez et al., 2007] Rodríguez, W.-R., Vaquero, C., Saz, O., and Lleida, E. (2007). Aplicación de las tecnologías del habla al desarrollo del prelenguaje y el lenguaje. In *Proceedings of the 2007 Congreso Latinoamericano de Ingeniería Biomédica (CLAIB)*, pages 1064–1067, Isla Margarita, Venezuela.

[Rodríguez et al., 2008b] Rodríguez, W.-R., Vaquero, C., Saz, O., and Lleida, E. (2008b). Speech technology applied to children with speech disorders. In *Proceedings of the 4th Kuala Lumpur International Conference on Biomedical Engineering*, pages 247–250, Kuala Lumpur, Malaysia.

[Rosenberg et al., 1992] Rosenberg, A., Delong, J., Lee, C., Juang, B., and Soong, F. (1992). The use of cohort normalized scores for speaker recognition. In *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP - Interspeech)*, Banff, Canada.

[Ruiz et al., 2008] Ruiz, I., Garcia, B., Mendez, A., and Mendezona, M. (2008). Oesophageal voice cycle detection in shimmer calculation algorithm. In *Proceedings of 7th WSEAS Conference on Signal Processing, Robotics and Automation (ISPRA)*, Cambridge, United Kingdom.

[Sanders et al., 2002] Sanders, E., Ruiter, M., Beijer, L., and Strik, H. (2002). Automatic recognition of Dutch dysarthric speech: A pilot study. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP-Interspeech)*, pages 661–664, Denver (CO), USA.

[Sawhney and Wheeler, 1999] Sawhney, N. and Wheeler, S. (1999). Using phonological context for improved recognition of dysarthric speech. Technical report, Speech Interface Group, MIT Media Lab.

[Saz et al., 2009a] Saz, O., Lleida, E., Abarca, L., and Mejuto, S. (2009a). Mouseclick: Acceso al ordenador a través de la voz. In *Proceedings of the IV Jornadas Iberoamericanas de Tecnologías de Apoyo a Discapacidad*, Madrid, Spain.

[Saz et al., 2009b] Saz, O., Lleida, E., Abarca, L., and Mejuto, S. (2009b). Mouseclick: Acceso al ordenador a través de la voz. In *Proceedings of the II Congreso Nacional de Comunicación Aumentativa*, Zaragoza, Spain.

[Saz et al., 2009c] Saz, O., Lleida, E., and Miguel, A. (2009c). Combination of acoustic and lexical speaker adaptation for disordered speech recognition. In *Proceedings of the 11th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 544–547, Brighton, United Kingdom.

[Saz et al., 2010a] Saz, O., Lleida, E., Rodríguez, V., Rodríguez, W.-R., and Vaquero, C. (2010a). The use of synthetic speech in language learning tools: Review and a case study. In Mullenix, J.-W. and Stern, D.-E., editors, *Computer Synthesize Speech Technologies: Tools for Aiding Impairment*. IGI Global Publishing, Hershey (PA), USA. In press.

[Saz et al., 2009d] Saz, O., Lleida, E., and Rodríguez, W.-R. (2009d). Acoustic Phonetic Decoding for assessment of mispronunciations in speakers with cognitive disorders. In *Proceedings of the 3rd Advanced Voice Function Assessment International Workshop (AVFA09)*, pages 129–132, Madrid, Spain.

[Saz et al., 2009e] Saz, O., Lleida, E., and Rodríguez, W.-R. (2009e). Avoiding speaker variability in pronunciation verification of children disordered speech. In *Proceedings of the 2009 Workshop on Children, Computer and Interaction*, Cambridge (MA), USA.

[Saz et al., 2006a] Saz, O., Miguel, A., Lleida, E., Ortega, A., and Buera, L. (2006a). Study of time and frequency variability in pathological speech and error reduction methods for Automatic Speech Recognition. In *Proceedings of the 2006 International Conference on Spoken Language Processing (ICSLP - Interspeech)*, pages 993–996, Pittsburgh (PA), USA.

[Saz et al., 2009f] Saz, O., Rodríguez, V., Lleida, E., Rodríguez, W.-R., and Vaquero, C. (2009f). An experience with a Spanish Second Language learning tool in a multilingual environment. In *Proceedings of the 2009 Workshop on Speech and Language Technologies in Education (SLaTE)*, Wroxall Abbey Estates, United Kingdom.

[Saz et al., 2010b] Saz, O., Rodríguez, V., Lleida, E., Rodríguez, W.-R., and Vaquero, C. (2010b). The use of multimodal tools for pronunciation training in second language learning of preadolescents. In Columbus, F., editor, *Language Teaching: Techniques, Developments and Effectiveness*. Nova Science Publishers, Hauppauge (NY), USA. Abstract accepted for publication.

[Saz et al., 2008a] Saz, O., Rodríguez, W.-R., Lleida, E., and Vaquero, C. (2008a). A novel corpus of children's impaired speech. In *Proceedings of the 2008 Workshop on Children, Computer and Interaction*, Chania, Greece.

[Saz et al., 2008b] Saz, O., Rodríguez, W.-R., Lleida, E., Vaquero, C., and Escartín, A. (2008b). Comunica - plataforma para el desarrollo, distribucin y evaluacin de herramientas logopdicas asistidas por ordenador. In *Proceedings of the V Jornadas en Tecnologías del Habla*, pages 37–40, Bilbao, Spain.

[Saz et al., ress] Saz, O., Rodríguez, W.-R., Vaquero, C., Escartín, A., Marcos, J.-M., and Canalís, C. (2009 (In press)). Consideraciones en el desarrollo de herramientas informáticas para logopedia en educación especial. *Maremagum - Publicación Galega sobre os Trastornos do Espectro Autista*, 13.

[Saz et al., 2008c] Saz, O., Simón, J., Dallal, E., Lleida, E., and Rose, R. (2008c). Modelado acústico de los errores en la realización de los fonemas para Reconocimiento Automático del Habla alterada. In *Proceedings of the IV Congreso de Fonética Experimental*, Granada, Spain.

[Saz et al., 2008d] Saz, O., Simón, J., Dallal, E., Lleida, E., and Rose, R. (2008d). Modelado acústico de los errores en la realización de los fonemas para Reconocimiento Automático del Habla alterada. *Language Design (Journal of Theoretical and Experimental Linguistics)*, 13:247–254.

[Saz et al., 2009g] Saz, O., Simón, J., Rodríguez, W.-R., Lleida, E., and Vaquero, C. (2009g). Analysis of acoustic features in speakers with cognitive disorders and speech impairments. *EURASIP Journal on Advances in Signal Processing*, Special Issue on Analysis and Signal Processing of Oesophageal and Pathological Voices.

[Saz et al., 2006b] Saz, O., Vaquero, C., Lleida, E., Marcos, J.-M., and Canalís, C. (2006b). Study of Maximum A Posteriori adaptation for Automatic Speech Recognition of pathological speech. In *Proceedings of the IV Jornadas en Tecnologías del Habla*, pages 395–398, Zaragoza, Spain.

[Saz et al., 2009h] Saz, O., Yin, S.-C., Lleida, E., Rose, R., Rodríguez, W.-R., and Vaquero, C. (2009h). Tools and technologies for computer-aided speech and language therapy. *Speech Communication*, 51(10):948–967.

[Sharma and Hasewaga-Johnson, 2009] Sharma, H.-V. and Hasewaga-Johnson, M. (2009). Universal access: Speech recognition for talkers with spastic dysarthria. In *Proceedings of the 11th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 1451–1454, Brighton, UK.

[Siohan et al., 2000] Siohan, O., Myrvoll, T.-A., and Lee, C.-H. (2000). Structural maximum a posteriori linear regression for fast HMM adaptation. In *Proceedings of the ASR2000 - Automatic Speech Recognition: Challenges for the new Millenium*, pages 120–127, Paris, France.

[Sprinthall, 2007] Sprinthall, R.-C. (2007). *Basic Statistical Analysis (8th Edition)*. Allyn & Bacon, Boston (MA), USA.

[Strik, 2001] Strik, H. (2001). Pronunciation adaptation at the lexical level. In *Proceedings of ISCA ITRW Workshop Adaptation Methods for Speech Recognition*, Sophia Antipolis, France.

[Szaszák et al., 2009] Szaszák, G., Sztahó, S., and Vicsi, K. (2009). Automatic intonation classification for speech training systems. In *Proceedings of the 11th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 1899–1902, Brighton, United Kingdom.

[Tang and Rose, 2007] Tang, Y. and Rose, R. (2007). Clustered maximum likelihood linear basis for rapid speaker adaptation. In *Proceedings of the 10th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 254–257, Antwerp, Belgium.

[Tepperman et al., 2006] Tepperman, J., Silva, J., Kazemzadeh, A., You, H., Lee, S., Alwan, A., and Narayanan, S. (2006). Pronunciation verification of children's speech for automatic literacy assessment. In *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP - Interspeech)*, Pittsburgh (PA), USA.

[Tsurutani et al., 2006] Tsurutani, C., Yamauchi, Y., Minematsu, N., Luo, D., Maruyama, K., and Hirose, K. (2006). Development of a program for self assessment of Japanese pronunciation by English learners. In *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP - Interspeech)*, Pittsburgh (PA), USA.

[van Doremalen et al., 2009] van Doremalen, J., Strik, H., and Cucchiarini, C. (2009). Utterance verification in language learning applications. In *Proceedings of the 2009 Workshop on Speech and Language Technologies in Education (SLaTE)*, Wroxall Abbey Estates, United Kingdom.

[Vaquero, 2006] Vaquero, C. (2006). Reconocedor de comandos orales para eliminar barreras de comunicación y movilidad en personas con discapacidades motrices y de comunicación. Proyecto Fin de Carrera, Departamento de Ingeniería Electrónica y Comunicaciones, University of Zaragoza, Zaragoza, Spain. Dirigido por O. Saz (Ponente E. Lleida).

[Vaquero et al., 2007] Vaquero, C., Saz, O., and Lleida, E. (2007). Tecnologías del habla para el desarrollo del lenguaje. In *Proceedings of the XVII Jornadas Telecom I+D*, Valencia, Spain.

[Vaquero et al., 2006] Vaquero, C., Saz, O., Lleida, E., Marcos, J.-M., and Canalís, C. (2006). Vocaliza: An application for computer-aided speech therapy in Spanish language. In *Proceedings of the IV Jornadas en Tecnologías del Habla*, pages 321–326, Zaragoza, Spain.

[Vaquero et al., 2008a] Vaquero, C., Saz, O., Lleida, E., and Rodríguez, W.-R. (2008a). E-inclusion technologies for the speech handicapped. In *Proceedings of the 2008 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4509–4512, Las Vegas (NV), USA.

[Vaquero et al., 2008b] Vaquero, C., Saz, O., Rodríguez, W.-R., and Lleida, E. (2008b). Human Language Technologies for speech therapy in Spanish language. In *Proceedings of the LangTech2008*, pages 129–132, Rome, Italy.

[Vicsi et al., 1999] Vicsi, K., Roach, P., Oester, A., Kacic, Z., Barczikay, P., and Sinka, I. (1999). SPECO: A multimedia multilingual teaching and training system for speech handicapped children. In *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 859–862, Budapest, Hungary.

[Wan and Carmichael, 2005] Wan, V. and Carmichael, J. (2005). Polynomial dynamic time warping kernel Support Vector Machines for dysarthric speech recognition with sparse training data. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 3321–3324, Lisbon, Portugal.

[Wik et al., 2009] Wik, P., Hincks, R., and Hirschberg, J. (2009). Responses to Ville: A virtual language teacher for Swedish. In *Proceedings of the 2009 Workshop on Speech and Language Technologies in Education (SLaTE)*, Wroxall Abbey Estates, United Kingdom.

[Witt and Young, 1997] Witt, S. and Young, S.-J. (1997). Computer-Assisted Pronunciation Teaching based on Automatic Speech Recognition. In *Proceedings of the International Conference on Language Teaching, Language Technology*, Groningen, The Netherlands.

[Witt and Young, 2000] Witt, S.-M. and Young, S.-J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2–3):95–108.

[Yin et al., 2008] Yin, S.-C., Rose, R., Saz, O., and Lleida, E. (2008). Verifying pronunciation accuracy from speakers with neuromuscular disorders. In *Proceedings of the 10th International Conference on Spoken Language Processing (ICSLP-Interspeech)*, pages 2218–2221, Brisbane, Australia.

[Yin et al., 2009] Yin, S.-C., Rose, R., Saz, O., and Lleida, E. (2009). A study of pronunciation verification in a speech therapy application. In *Proceedings of the 2009 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4609–4612, Taipei, Taiwan.

[Young et al., 2006] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, Cambridge, UK.

[Yu et al., 2001] Yu, P., Ouaknine, M., Revis, J., and Giovanni, A. (2001). Objective voice analysis for dysphonic patients: A multiparametric protocol including acoustic and aerodynamic measurements. *Journal of Voice*, 15(4):529–542.