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

Oscar Saz

Communications Technology Group (GTC)
Aragón Institute for Engineering Research (I3A)
University of Zaragoza, Spain
oskarsaz@unizar.es

Abstract

This thesis deals with the research and development of speech technology-based systems for the requirements of users with different impairments and disabilities, with the final aim of improving their quality of life. Speech disorders are shown to be a major challenge in the work with these users. This work performs all the steps in the research in speech technologies: starting with the acquisition of an oral corpus from young impaired speakers, the analysis of the acoustic and lexical variations in the disordered speech and the characterization of speaker dependent Automatic Speech Recognition (ASR) systems adapted to the acoustic and lexical variants introduced by these speakers. Furthermore, automated methods for detection and correction of lexical mispronunciations are also evaluated. The results of the experiments show the on-going possibility for developing a fully personalized ASR system for handicapped users that learns the speaker's speech characteristics on-line: while the user interacts with the recognition system. The development of speech therapy tools based on the knowledge gained is another outcome of the present thesis, where the development of "Comunica" aims to improve the possibilities for semi-automated speech therapy in Spanish.
Index Terms: speech disorders, speaker personalization, language learning

1. Introduction

Communication disorders are a heavy limitation for those who suffer them, and the impossibility to communicate with others impedes the social inclusion and development of impaired people. Difficulties in the access to education, impossibility in the access to work or social exclusion are some of the consequences of communicative impairments.

New technologies can facilitate communication and knowledge as they make access to information universal, immediate and ubiquitous. Unfortunately, current interfaces based on peripherals like mouse or keyboard are not accessible to disabled users as they require good levels of motor control and cognitive capabilities. Research on new forms on human computer interfaces towards more natural, adaptable and accessible interaction is currently being done in many facets like eye gaze tracking, movement tracking, brain computer interface or speech, which is the subject of this thesis.

1.1. Speech Technology for the Handicapped

Different systems based on Automatic Speech Recognition (ASR) and Text-To-Speech (TTS) synthesis have already been researched for their use by the handicapped community. The STARDUST project aimed to provide oral commands and control of a home environment [1, 2]; the Vocal Joystick was designed to provide accessibility to computers by heavily impaired users [3, 4]; and, finally, the VIVOCA project created communicative aids able to recreate the speech from a disordered user [5, 6].

Education is another area of work where speech technologies can help the handicapped community; Computer-Aided Speech and Language Therapy (CASLT) tools, a subset within the broader domain of Computer-Aided Language Learning (CALL) tools [7], bring speech therapy to individuals with communication difficulties [8, 9].

One of the main difficulties for the research in these lines of work is the lack of resources (corpora, databases) to characterize and study the main features of disordered speech and to develop and evaluate automated speech systems. Corpora like the Whitaker [10] and Nemours [11] databases were collected at the early stages of the interest in disordered speech. Currently, the Universal Access Database [12] is the largest database covering the domain of dysarthric and disordered speech.

1.2. Motivation and Objectives

This thesis has grown up thanks to the interest of different people and institutions in the creation of software and systems based in speech technologies for education, inclusion and assistance purposes. These institutions like the Public School for Special Education (CPEE) "Alborada", CADIS-Huesca, ASPACE-Huesca or the Vienna International School (VIS) are the great motivators of this thesis.

The objectives of this thesis cover two different aspects in the technological work. From a scientific point of view, it aims to provide the community with a fully functional corpus of speech disorders in Spanish, to learn and study different adaptation techniques to these disorders in ASR systems, and to develop techniques for the assessment of speech proficiency. From a practical point of view, it aims to develop devices for control and interaction based on speech and voice which are accessible for handicapped users and CALL tools for students with different needs in learning communication through language.

2. Methodology and Corpus

The methodology of work was as presented in Figure 1. The corpus collected with different impaired voices served as the foundations of the work. Posteriorly, a main line of work appeared towards a better knowledge of the properties of speech

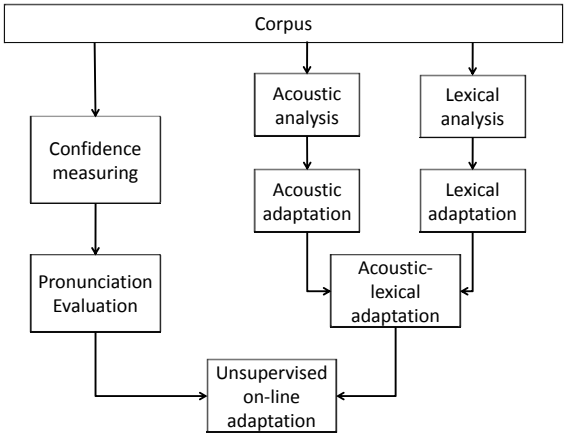

Figure 1: *Methodology of work.*

impairments through different analyses. These led to proposals for speaker adaptation in the acoustic and lexical models of the ASR, which were merged in acoustic-lexical speaker adaptation. A parallel line studying how to evaluate correctly the pronunciation of the impaired speakers and transform this in confidence measures was finally merged with the speaker adaptation into a proposal for unsupervised on-line personalized system.

2.1. The "Alborada-I3A" corpus

The "Alborada-I3A" corpus was recorded with the intention of filling the gap in resources and databases of disordered speech in Spanish [13]. It was recorded in collaboration with different educative institutions and is fully available for research.

The core of the corpus were 14 young speakers suffering different physical and cognitive handicaps and also suffering very different speech and language disorders. These speakers were 7 boys and 7 girls from 11 to 21 years old. A set of 4 isolated word sessions were recorded from each speaker, with a vocabulary of 57 words per session, for a total of 3,192 isolated word utterances. The set of words used, the Induced Phonological Register (RFI) [14], is well known among speech therapists in Spanish and contains a all phonemes of Spanish language in different positions and contexts. Furthermore, 232 young unimpaired speakers in the same age range than the impaired speakers (10-18 years old) were later recorded to model and characterize the speech in children and young adults with a proper correct speech.

In order to characterize phonologically the production of errors in the impaired speakers, a labeling was carried out by a set of experts to determine whether each phoneme in the pronunciation of the speakers was correctly pronounced, mistaken or substituted, or deleted by the speaker. The results showed a big relevance of the disorders, as around 10% of the phonemes were substituted and 7% were deleted.

The baseline ASR experiments with this corpus pointed out the big influence of the disorders on the performance of the system. While the unimpaired speakers were in a 4% of Word Error Rate (WER) on adult speech models, the impaired peers reached 37% WER. Task dependent models, retrained on the 232 unimpaired speakers allowed for a reduction of the WER to 28%, which was isolated as been due to acoustic and lexical disorders.

3. Analyses of Disordered Speech

Before starting any experimentation with the corpus, several studies were carried out to understand how the disorders affected all the facets of speech production. These studies were expected to be relevant on how to face all the problematic regarding recognition and evaluation of disordered speech.

3.1. Acoustic Analysis

A study on vowel production by the impaired speakers was made to find any differences between unimpaired and impaired speakers. The features which were under study were: First two formant frequencies (F_1 and F_2), fundamental frequency value (F_0), intensity value and duration. Speech processing methods like Linear Predictive Coding (LPC) and autocorrelation were used to calculate these features.

A degradation of the quality of vowel production was measured in some of these features, showing the inability of the impaired speakers for a precise control of articulation. This degradation affected in the following ways: Reduction of distance between vowels in the formant map, especially between /a/, /e/ and /o/; loss of distinction in the production of intensity between stressed and unstressed vowels; and a higher variability in vowels duration [15].

3.2. Lexical Analysis

The patterns of production of mistakes in the speakers were studied from a phonological point of view. A first approach did not show an influence of the phonological properties of phonemes (point and manner of articulation). On the contrary, the context and position of the phoneme in the syllable were the most prominent factors in mispronunciations. The main findings of this analysis were the significant reductions of vowels in diphthongs, consonants in coda position and consonant clusters.

The comparison of the pronunciation patterns of this impaired speakers with the pattern in young children in learning stages (3-6 years old) showed how young impaired speakers presented similar errors to small children. This pointed out that the possible origin of this language disorders might be a delay in language acquisition due to the cognitive delays of these speakers.

4. Techniques for Personalization

Personalization is a key point when developing speech interfaces for the handicapped. The influence of the specific impairment of each user is so unique that they require that the system is fully adaptive to the user instead of the user adapting to the system.

4.1. Speaker Adaptation

The studies carried out in the effects of speech disorders have shown the existence of acoustic distortion in the speakers' voice as well as phonological and lexical disorders leading to mispronunciations at this level [16]. Adaptation was proposed at three levels:

Acoustic adaptation for these speakers can provide a better matching of acoustic models to their speech properties. These properties not only include inter speaker variability as for the rest of the population, but the modeling of their speech disorders and how they modify the acoustic production of speech. Lexical adaptation provides a correct modeling of the pronunciations that these speakers are uttering, characterized by a large number

of substitutions and deletions at the phoneme level. While lexical adaptation is usually not required for unimpaired speakers, these speakers might find relevant benefits with it. Acoustic-lexical adaptation merges the two previous approaches, but it is important to understand the correlations between them as they provide different ways to model similar effects of speech.

The proposal for acoustic adaptation was Maximum A Posteriori [17] and Maximum Likelihood Linear Regression (MLLR) [18], while the proposal for lexical adaptation was a data-driven approach, which learned the speaker's transcriptions through Acoustic Phonetic Decoding (APD). After several experiments, the joint use of acoustic and lexical adaptation produced a relative improvement in the WER of 46%.

4.2. Pronunciation Evaluation

Pronunciation verification is a relevant issue in the work with speech disorders, it can serve to correct the speaker's pronunciation or to discard incorrect pronunciations in speech systems. Among several other proposals [19, 20], the work focused on normalization of the phoneme scores to different combinations of competing phonemes in the same way that the test normalization (or t-norm) method for speaker verification tasks.

The results showed that a correct selection of the competing phonemes [21] could achieve better results in terms of Equal Error Rate (EER) in the pronunciation verification tasks than known techniques like the Goodness of Pronunciation (GOP) [22]. GOP can also be seen as normalization technique where all the competing phonemes are considered. Better result of EER was lowered to 16%.

4.3. Proposal of On-line Unsupervised Personalized System

Finally, a proposal for on-line unsupervised personalization was provided in Figure 2. In this system, the user is fully unaware of the procedure in which the personalization is performed constantly in the underlying loop. During the use of the ASR system, the confidence measuring algorithm discard those transcriptions which might be inaccurate (either due to recognition mistakes or due to pronunciation mispronunciations) and stores in a buffer all the signals which are considered useful for a posterior adaptation. Once the buffer has sufficient amount of reliable data, speaker adaptation is performed according to the transcriptions obtained by the ASR system validated by the confidence measure. The new models (acoustic and/or lexical) are inserted in the ASR system to improve the recognition accuracy of the user's speech. This process can be repeated to keep providing further adaptation to the speaker.

After some preliminary experiments in this proposal, it was seen how an iterative procedure of adaptation like the one depicted in Figure 2 could outperform in terms of improvement of the WER a similar proposal using the same amount of adaptation data in one single stage. Several concerns have to be taken into account in this proposal, especially the influence of the initial ASR stage to obtain the transcriptions and the accuracy of the confidence measure algorithm.

5. Speech Based Applications

The experimental work in the thesis has been accompanied by work on the development and deployment of speech-based tools to improve the quality of life of the handicapped. "VozClick", developed for ASPACE-Huesca, aims to substitute physical switches for severely handicapped people in their access to computer applications [23]. It transforms a pulse of vocal emission by the user into a event in the computer substituting peripherals like mouse or keyboard.

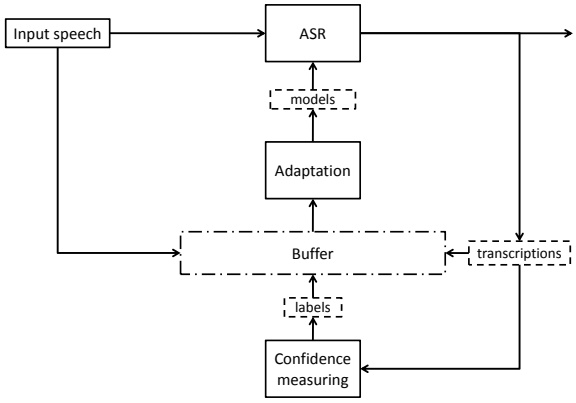

Figure 2: *On-line personalization.*

It has been in the development of CALL tools where greater results have been achieved during this work where "Comunica" has been a framework on the research and development of CASLT tools in Spanish [24]. "Comunica" consists of "PreLingua" for the training of phonatory skills in small handicapped children [25], "Vocaliza" for the training of articulation skills [26] and "Cuéntame" for linguistic skills. Their open and free distribution through the Internet[1] has shown the great interest of speech therapists in Spain and Latin America for this kind of technical aids and their appreciation for the tools in "Comunica"

The results in pronunciation verification achieved in the thesis were introduced in "Vocaliza" to provide a tool for the training of Spanish as a Second Language (L2) to children. The results of an experience carried out in this field showed the usefulness of computer-aided tools with a dedicated interface and a correct use of speech technologies [27, 28, 29].

6. Conclusions

This thesis has supposed a relevant effort in all the objectives proposed at the beginning of the work. The corpus has showed to be useful and it has put the interest of the community on this specific task. Different personalization techniques have been evaluated and the performance of them has been framed. Confidence measuring and pronunciation assessment has shown significant improvements and solid results.

The thesis has also discussed several subjects regarding the origins of the speech disorders, their affection in the speech production of the users and their effect on the performance of automated speech recognition and assessment systems. More precisely, disorders at the lexical level have had a special treatment in their analysis and evaluation in the thesis, compared to previous works focusing mostly on the acoustic side of speech.

Finally, further work which has arisen from the thesis has to be oriented towards the study on personalization techniques which can work in cases of unsupervised data and data sparsity. Techniques which take into account the mutual information of different sources might be useful for further improvement in confidence measuring tasks. Finally, all this work still has open road to be introduced in real systems to provide accessibility and inclusion for handicapped people.

[1] http://www.vocaliza.es

7. Acknowledgments

This thesis was supported by the FPU program of the Spanish Ministry of Education. Part of the work was funded under national projects TIN 2005-08660-C04-01 and TIN2008-06856-C05-04. The thesis was carried out under the supervision of Prof. Eduardo Lleida at the Department of Electrical Engineering and Communications of the University of Zaragoza (Spain). Work was shared with Prof. Phil Green and Dr. Stuart Cunningham at the Department of Computer Science at the University of Sheffield (United Kingdom) in June-September 2006 and with Prof. Richard Rose at the Department of Electrical and Computer Engineering at McGill University (Montreal, Canada) in June-September 2007.

8. References

[1] P. Green, J. Carmichael, A. Hatzis, P. Enderby, M. Hawley, and M. Parker, "Automatic Speech Recognition with sparse training data for dysarthric speakers," in *Proceedings of the 8th Eurospeech*, Geneva, Switzerland, September 2003, pp. 1189–1192.

[2] M. Hawley, P. Enderby, P. Green, S. Brownsell, A. Hatzis, M. Parker, J. Carmichael, S. Cunningham, P. O'Neill, and R. Palmer, "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, August 2003.

[3] J. Bilmes, J. Malkin, X. Li, S. Harada, K. Kilanski, K. Kirchhoff, R. Wright, A. Subramanya, J. Landay, P. Dowden, and H. Chizeck, "The vocal joystick," in *Proceedings of the 2006 ICASSP*, Toulouse, France, May 2006, pp. 625–628.

[4] S. Harada, J. Landay, J. Malkin, X. Li, and J. Bilmes, "The Vocal Joystick: evaluation of voice-based cursor control techniques for assistive technology," *Disability and Rehabilitation: Assistive Technology*, vol. 3, no. 1, pp. 22–34, 2008.

[5] S. Creer, S.-P. Cunningham, P.-D. Green, and K. Fatema, "Personalizing synthetic voices for people with progressive speech disorders: Judging voice similarity," in *Proceedings of the 11th Eurospeech*, Brighton, UK, September 2009, pp. 1427–1430.

[6] S. Creer, P. Green, S. Cunningham, and J. Yamaghisi, "Building personalized synthesized voices for individuals with dysarthria using the HTK toolkit," in *Computer Synthesized Speech Technologies: Tools for Aiding Impairment*, J.-W. Mullenix and S.-E. Stern, Eds. Hershey (PA), USA: IGI Publishing, 2010.

[7] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[8] K. Vicsi, P. Roach, A. Oester, Z. Kacic, P. Barczikay, and I. Sinka, "SPECO: A multimedia multilingual teaching and training system for speech handicapped children," in *Proceedings of the 6th Eurospeech*, Budapest, Hungary, September 1999, pp. 859–862.

[9] A.-M. Oester, D. House, A. Protopapas, and A. Hatzis, "Presentation of a new EU project for speech therapy: OLP (Ortho-Logo-Paedia)," in *Proceedings of the XV Swedish Phonetics Conference (Fonetik 2002)*, Stockholm, Sweden, May 2002, pp. 45–48.

[10] J.-R. Deller, M.-S. Liu, L.-J. Ferrier, and P. Robichaud, "The whitaker database of dysarthric (cerebral palsy) speech," *Journal of the Acoustical Society of America*, vol. 93, no. 6, pp. 3516–3518, 1993.

[11] X. Menéndez-Pidal, J.-B. Polikoff, S.-M. Peters, J. Lorenzo, and H.-T. Bunnell, "The Nemours database of dysarthric speech," in *Proceedings of the 4th ICSLP*, Philadelphia (PA), USA, October 1996, pp. 1962–1965.

[12] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. France, "Dysarthric speech database for universal access research," in *Proceedings of the 10th ICSLP*, Brisbane, Australia, September 2008, pp. 1741–1744.

[13] O. Saz, W.-R. Rodríguez, E. Lleida, and C. Vaquero, "A novel corpus of children's impaired speech," in *Proceedings of the 2008 Workshop on Children, Computer and Interaction*, Chania, Greece, October 2008.

[14] M. Monfort and A. Juárez-Sánchez, *Registro Fonológico Inducido (Tarjetas Gráficas)*. Madrid, Spain: Ed. Cepe, 1989.

[15] O. Saz, J. Simón, W.-R. Rodríguez, E. Lleida, and C. Vaquero, "Analysis of acoustic features in speakers with cognitive disorders and speech impairments," *EURASIP Journal on Advances in Signal Processing*, vol. Special Issue on Analysis and Signal Processing of Oesophageal and Pathological Voices, 2009.

[16] O. Saz, E. Lleida, and A. Miguel, "Combination of acoustic and lexical speaker adaptation for disordered speech recognition," in *Proceedings of the 11th Eurospeech*, Brighton, United Kingdom, September 2009, pp. 544–547.

[17] J.-L. Gauvain and C.-H. Lee, "Maximum A Posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[18] C.-J. Legetter and P.-C. Woodland, "Maximum Likelihood Linear Regression for speaker adaptation of the parameters of continous density Hidden Markov Models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[19] S.-C. Yin, R. Rose, O. Saz, and E. Lleida, "Verifying pronunciation accuracy from speakers with neuromuscular disorders," in *Proceedings of the 10th ICSLP*, Brisbane, Australia, September 2008, pp. 2218–2221.

[20] ——, "A study of pronunciation verification in a speech therapy application," in *Proceedings of the 2009 ICASSP*, Taipei, Taiwan, April 2009, pp. 4609–4612.

[21] O. Saz, E. Lleida, and W.-R. Rodríguez, "Avoiding speaker variability in pronunciation verification of children disordered speech," in *Proceedings of the 2009 Workshop on Children, Computer and Interaction*, Cambridge (MA), USA, November 2009.

[22] S.-M. Witt and S.-J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2–3, pp. 95–108, 2000.

[23] O. Saz, E. Lleida, L. Abarca, and S. Mejuto, "Mouseclick: Acceso al ordenador a través de la voz," in *Proceedings of the IV Jornadas Iberoamericanas de Tecnologías de Apoyo a Discapacidad*, Madrid, Spain, October 2009.

[24] O. Saz, S.-C. Yin, E. Lleida, R. Rose, W.-R. Rodríguez, and C. Vaquero, "Tools and technologies for computer-aided speech and language therapy," *Speech Communication*, vol. 51, no. 10, pp. 948–967, 2009.

[25] W.-R. Rodríguez, C. Vaquero, O. Saz, and E. Lleida, "Speech technology applied to children with speech disorders," in *Proceedings of the 4th Kuala Lumpur International Conference on Biomedical Engineering*, Kuala Lumpur, Malaysia, June 2008, pp. 247–250.

[26] C. Vaquero, O. Saz, E. Lleida, and W.-R. Rodríguez, "E-inclusion technologies for the speech handicapped," in *Proceedings of the 2008 ICASSP*, Las Vegas (NV), USA, April 2008, pp. 4509–4512.

[27] O. Saz, V. Rodríguez, E. Lleida, W.-R. Rodríguez, and C. Vaquero, "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, 2009.

[28] O. Saz, E. Lleida, V. Rodríguez, W.-R. Rodríguez, and C. Vaquero, "The use of synthetic speech in language learning tools: Review and a case study," in *Computer Synthesize Speech Technologies: Tools for Aiding Impairment*, J.-W. Mullenix and D.-E. Stern, Eds. Hershey (PA), USA: IGI Global Publishing, 2010.

[29] O. Saz, V. Rodríguez, E. Lleida, W.-R. Rodríguez, and C. Vaquero, "The use of multimodal tools for pronunciation training in second language learning of preadolescents," in *Language Teaching: Techniques, Developments and Effectiveness*, F. Columbus, Ed. Hauppauge (NY), USA: Nova Science Publishers, 2010.