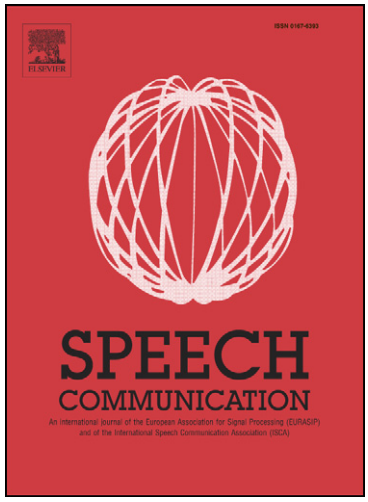


Tools and technologies for Computer-Aided Speech and Language Therapy

Oscar Saz, Shou-Chun Yin, Eduardo Lleida, Richard Rose, Carlos Vaquero, William R. Rodríguez

PII: S0167-6393(09)00066-1
DOI: 10.1016/j.specom.2009.04.006
Reference: SPECOM 1801

To appear in: *Speech Communication*

Received Date: 1 July 2008
Revised Date: 31 March 2009
Accepted Date: 9 April 2009

Please cite this article as: Saz, O., Yin, S-C., Lleida, E., Rose, R., Vaquero, C., Rodríguez, W.R., Tools and technologies for Computer-Aided Speech and Language Therapy, *Speech Communication* (2009), doi: 10.1016/j.specom.2009.04.006

# Tools and Technologies for Computer-Aided Speech and Language Therapy

Oscar Saz [a,1], Shou-Chun Yin [b,2], Eduardo Lleida [a], Richard Rose [b], Carlos Vaquero [a] and William R. Rodríguez [a]

[a] *Communications Technology Group (GTC), Aragón Institute for Engineering Research (I3A), University of Zaragoza, María de Luna 1. 50018, Zaragoza, Spain.*
*{oskarsaz,lleida,cvaquero,wricardo}@unizar.es*
*Phone (+34) 976 762705, Fax: (+34) 976 762111*

[b] *Department of Electrical and Computer Engineering, McGill University, 3480 University Street, Montréal (QC) H3A 2A7, Canada.*
*shou-chun.yin@mail.mcgill.ca,rose@ece.mcgill.ca*
*Phone (514) 398 1749, Fax: (514) 398 4470*

**Abstract**

This paper addresses the problem of Computer-Aided Speech and Language Therapy (CASLT). The goal of the work described in the paper is to develop and evaluate a semi-automated system for providing interactive speech therapy to the increasing population of impaired individuals and help professional speech therapists. A discussion on the development and evaluation of a set of interactive therapy tools, along with the underlying speech technologies that support these tools is provided. The interactive tools are designed to facilitate the acquisition of language skills in the areas of basic phonatory skills, phonetic articulation and language understanding primarily for children with neuromuscular disorders like dysarthria. Human-machine interaction for all of these areas requires the existence of speech analysis, speech recognition, and speech verification algorithms that are robust with respect to the sources of speech variability that are characteristic of this population of speakers. The paper will present an experimental study that demonstrates the effectiveness of an interactive system for eliciting speech from a population of impaired children and young speakers ranging in age from 11 to 21 years. The performance of automatic speech recognition (ASR) systems and subword-based pronunciation verification (PV) on this domain are also presented. The results indicate that ASR and PV systems configured from speech utterances taken from the impaired speech domain can provide adequate performance, similar to the experts' agreement rate, for supporting the presented CASLT applications.

*Key words:* Spoken Language Learning, Speech Disorders, Speech Corpora, Automatic Speech Recognition, Pronunciation Verification

## 1 Introduction

It is often the case that there are insufficient resources to provide for the acquisition of speech and language skills for handicapped children and young adults suffering from speech disorders. These handicaps might as well be physical or developmental disorders leading to dysarthria, a general type of speech impairment caused by a disorder of the neuromuscular system. This could be generated by a developmental disability like Down's Syndrome or by cerebral palsy caused by stroke, as well as by organic disorders in the phonatory, articulatory or auditory systems. Speech therapy for these individuals generally involves extended interaction between an individual student and a skilled therapist. As a result, the time and the expense of providing this therapy for all impaired students can make it impractical to serve a large student population.

This paper describes interactive tools that were developed to reduce the time required and also perhaps to reduce the level of expertise required from the therapist for providing the interactive component of the therapy. There are three general areas of diagnosis and treatment for speech and language disorders that are reviewed in Section 2. The first is the acquisition of basic phonatory skills. These skills include control of basic functions like voicing, speech intensity, breathing and tone. Automated interactive applications have been implemented for developing these skills, but are not evaluated in this paper. The second is acquisition of the phonetic system of a language, this is, the pronunciation of the set of sounds of the language and the way they create syllables and words. The user interaction in this case involves the patient receiving feedback evaluating quality of pronunciation of words presented in a therapy session. Automatic procedures for verifying the quality of pronunciations are proposed and evaluated in Section 6 to provide this feedback. The last area of diagnosis and treatment is language understanding. In this case, the user interaction involves interactive dialogs with the student to evaluate skills in various question-answering scenarios. Automatic speech recognition (ASR) performance is evaluated under several adaptation paradigms in Section 5 using utterances collected from a population of impaired children and young adults to determine whether adequate performance (comparable to the agreement of a set experts that have labeled the same data) can be obtained

[1] This work was supported under TIN-2005-08660-C04-01 from MEC of the Spanish government

[2] This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Program Number 307188-2004

for operation in this application domain.

A study was performed to characterize the semi-automated interactive speech therapy domain. The goal was to evaluate the feasibility of a more efficient and lower cost methodology for diagnosis and treatment of students with neuromuscular disorders that could help human therapists. This methodology includes, at its lowest level, an interactive dialog between the student and an automated system for performing training, collecting speech from the student, and providing performance feedback. At the next level, a mechanism exists for a non-expert to measure the performance of the student. This was provided through a simple, easily reproducible scheme for labeling utterances at the phonemic level according to the accuracy of pronunciation. At the highest level, the speech therapist can assimilate the evaluations obtained from the students interactive sessions, review sample utterances, provide performance assessment, and prescribe additional therapy for the student.

The study was performed using a speech corpus containing utterances elicited from a population of impaired children and young speakers ranging in age from 11 to 21 years. These speakers were prompted using on-screen text and images to utter words from a phonetically balanced vocabulary designed for diagnosis of speech impairments (Monfort and Juárez-Sánchez, 1989). In the end, the corpus included 14 impaired speakers as well as 168 unimpaired speakers. The study had two components. The first component was to determine whether natural speech could be elicited using this paradigm and to determine if non-experts could reliably specify where portions of utterances have been mispronounced. This issue was addressed by implementing and evaluating a labeling scheme on this corpus where non-expert human labelers provided phoneme-level labels to indicate phoneme mispronunciations and deletions. The second component of the study was the use of this corpus to develop and evaluate ASR and Pronunciation Verification (PV) techniques as discussed above.

There has been a great deal of research in the area of Computer-Aided Speech and Language Therapy (CASLT) over the last decade and research on various aspects of speech interfaces for human-machine interaction has been performed before (Oester et al., 2002; Vicsi et al., 1999; García-Gómez et al., 1999). The work described in this paper extends this work by considering the components of the proposed multi-level methodology proposed above for diagnosis and treatment of impaired children. The paper begins in Section 2 by reviewing the procedures involved in traditional speech and language therapy and in novel CASLT applications. In Section 3, the components of the semi-automated "Comunica" system for human-computer interaction practices in speech and language therapy, as they relate to the procedures performed in traditional therapy, are described. A characterization of the human-machine interaction in articulation training is provided in the form of the speech corpus

and phoneme level transcription scheme in Section 4. Finally, the ASR and PV techniques are presented and evaluated for this task domain in Sections 5 and 6. Discussion and conclusions are extracted in Sections 7 and 8.

## 2 Speech and Language Disorders and Therapy

Traditional methods for speech and language therapy are based on the direct interaction between patient and therapist via a set of activities developed by the therapist for the diagnosis and treatment of the disorders of the patient. This direct interaction is necessary and effective in giving a personal feedback to every patient. But this way of working would require a high number of therapists to help all the possible patients in a school or institution without slowing down their possibilities of maximum progress, that unfortunately is not feasible now in most of the cases. Furthermore, this interaction is entirely based on the subjective evaluation of the therapist; hence, this evaluation might vary along time as the therapist gets used to the patient's speech or when the patient changes of therapist for whatever reason. The development of CASLT systems may overcome these two drawbacks and help professional therapists by providing a semi-automated way for speech therapy (several patients can work at the same time) and with an evaluation whose change over time will only reflect the changes in the patient's speech. This Section brings a review on traditional speech and language therapy techniques and recent speech technologies-based techniques.

### *2.1 Speech acquisition - Phonation*

Prior to language acquisition, children need to learn to control their own speech production with skills like breathing, control of tone and intensity and vocalization. Developmental disabilities and neuromuscular disorders can create a delay in the acquisition of these abilities that make the children unable to start articulating the first sounds and words. Speech therapists have been making use of game-like activities to train these skills: Moving the fire of a candle will help an impaired child to learn the control of breathing; vocalization in front of a mirror will help the therapist in providing feedback about the correct articulation of the vowels. Following these ideas, several handbooks (Acero-Villán et al, 2005) are published to discuss these strategies and propose new ideas.

Section 3.1 will present the approach in this work to computer-aided phonation acquisition in "PreLingua". This tool aims to translate to a computer-based framework a set of activities in which the patient trains all these abilities

mentioned above in a direct and simple way.

### *2.2 Language acquisition - Articulation*

The main step of speech and language acquisition under study in this work is the acquisition of the phonetic system of the language. Speech and language therapy in this level consists on sessions between the speech therapist and the patient who utters different words and receives an evaluation on the correctness of the pronunciation by the therapist. Books providing different set of especially chosen words are often published (Monfort and Juárez-Sánchez, 1989; Albor, 1991; Aguinaga et al., 2004) to help the community of speech therapists in their work.

Section 3.2 will introduce "Vocaliza", the proposal in this work for articulation acquisition and speech training. This application aims to train the phonological abilities of children while also introducing to them the semantics and syntax of the language via different activities. Furthermore, the application relies on Augmentative and Alternative Communication (AAC) systems to make the work friendlier for the patients and easier for the professionals in speech and language therapy. The technologies used to provide language improvement are ASR and PV. Sections 5 and 6 will measure the performance of these technologies to achieve the proposed goals.

### *2.3 Language acquisition - Understanding*

The final step in the language acquisition is the ability to use it as a tool to fulfill the daily situations. When children master the phonological level of the language (articulation) they have to start using this language to interact with the world. One important task is to be able to describe their environment or answering questions about it (descriptive skills) and it is also important to establish dialogs in a real environment to achieve their desired objectives in their daily life (dialog skills).

This area of speech and language therapy also requires knowledge in psychology to be able to motivate the patient to achieve goals and objectives in the everyday life through language. So usually, the speech training consists on the proposal of an activity with images and text to be solved via dialog with the educator (that is, answering to a given question, the description of a given situation or progressing in a simulated scenario) as proposed in several handbooks (Monfort and Monfort-Juárez, 2001).

Section 3.3 will introduce "Cuéntame", an application that intends to motivate

children to use language as a descriptive and interactive tool. "Cuéntame" presents a set of oral scenarios that the user has to solve in order to obtain an audio-visual reward. AAC systems are used in a similar way as in "Vocaliza" to make the tool accessible to children with diverse sensory needs (hearing-impaired, vision-impaired,...). ASR is again the basis of the application in the interactive dialogs.

### *2.4 A review of CASLT applications*

The development of CASLT software has been a major issue since the 1990's. Speech technologies have been increasing their robustness to different environment conditions and this has allowed the creation of new tools as it will be reviewed in this Section.

Considering phonatory skills in young children, the increase in the computation power of new computers made possible the analysis of features such as pitch or formants in real time applications for home computers. IBM's SpeechViewer was a very popular commercial approach for the training of speech skills as explained in speech therapy in Section 2.1. This application used a very simple speech interface of games, in which the ability of the user was tested in producing speech with varying intensity, pitch or formants. This software was a very interesting and effective work in improving young children's speech skills (Pratt et al., 1993) with a good approach in creating the interface with the patients.

In terms of language training in the articulation level, several research projects have explored the possibilities of providing this level of language therapy. It is to mention the special interest that the 6th Framework Program of the European Union ("Quality of Life and Management of Life Resources") had in this issue. All these approaches deal with important issues in the development of CASLT applications like the user interface or the feedback provided to the user. The approach of works like Optical-Logo-Therapy (Hatzis, 1999) and its successor Orto-Logo-Paedia (Oester et al., 2002) was to give the user a visual feedback in the correct positioning of the articulatory elements (tongue, palate, teeth) to produce correctly the different sounds. HARP (Lefévre, 1996) focused on the special speech therapy needs of the hearing impaired community, considering that these individuals can not have an audio feedback on the correct pronunciation of sounds. Other projects like SPECO (Vicsi et al., 1999) were intended as a multilingual speech therapy software for several European languages based on an audiovisual interaction scenario with game-like activities. The use of AAC systems has been a priority for other tools (Granstroem, 2005), while some works aimed for the speech rehabilitation of patients having suffered a laryngectomy (Kornilov, 2004) have been also reviewed.

Finally, there are two major fields in these applications which are gaining importance nowadays. First, the development of reading tutors for the assessment of the literacy of young children. The research in these applications is rising as the possibility of including Large Vocabulary Continous Speech Recognition (LVCSR) in real time applications is a matter of fact today. Recent approaches include (Duchateau et al., 2007) and (Gerosa and Narayanan, 2008) in which the user can approach the system with natural language to assess the comprehension of a text via open-ended questions. The development of Second Language (L2) learning tools is also an incredibly increasing line today due to the need of achieving a good assessment in foreign languages. This interest is growing due to the need of a correct use of foreign languages in the world of business and academia. In L2 learning, tools are usually aimed either to the acquisition of the sounds and phonetics of the target language (Cucchiarini et al., 2007), or to the acquisition of the correct grammar or syntax of the target language (Ito et al., 2008). These grammatical and syntax approaches can be shared with the development of CASLT tools for the training of the language understanding abilities in disabled speakers.

## 3 Tools for speech and language therapy: "Comunica"

The development of tools for speech and language therapy requires the work of an interdisciplinary team that merge expertise in the field of speech technologies and in the field of speech and language therapy and education. Thus, this work was supported by the staff and educators of the Public School for Special Education (CPEE) "Alborada", located in Zaragoza (Spain). Their collaboration in other projects has been fruitful in the fields of AAC technologies like communication boards, computer handling aids and space-time orientation devices (Martínez et al., 2007).

Three tools are part of the "Comunica" project (Escartín et al., 2008): "PreLingua" teaches basic phonation skills to children with neuromuscular disorders. "Vocaliza" (the firstly developed tool in this framework) aims to train mainly the articulatory level of language. Finally, "Cuéntame" attempts to introduce impaired children population to language understanding.

"Comunica" is an effort of the researchers in speech technologies of the Aragón Institute for Engineering Research (I3A) with the supervision of the CPEE "Alborada". The Aragonese Center in Education Technologies (CATEDU) is currently funding the creation of free pictorial items in use in several technical aids like "Comunica".

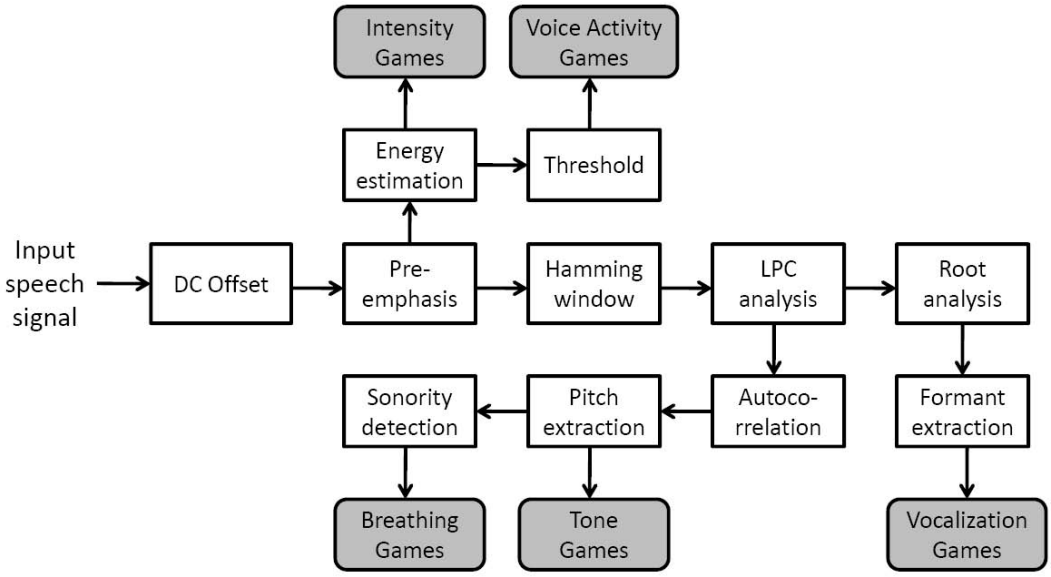


Fig. 1. Feature extraction in "PreLingua"

### *3.1 "PreLingua"*

"PreLingua" gathers a set of game-like applications that use speech processing to train children with speech developmental delays, aiming to help the work in speech therapy oriented to phonation as seen in Section 2.1. A feature extraction diagram like the one shown in Figure 1 is used for the training of five speech skills in the games (voice activity, intensity, breathing, tone and vocalization). After signal preprocessing (DC offset and pre-emphasis), energy of each frame is calculated (used in intensity games) and a threshold is applied to determine if voice is presented on the frame (used for voice activity games). Signal is then windowed and a Linear Prediction Coefficients (LPC) (Rabiner and Schafer, 1978) analysis applied to estimate the vocal tract transfer function. From this point, the autocorrelation of the prediction error leads to the extraction of the fundamental frequency value (used in tone games) and sonority level (used in breathing games); and the extraction of the roots of the transfer function is used to extract the formant frequencies values (used in vocalization games).

*Voice activity games* (Rodríguez et al., 2008) are oriented to children with a developmental disability that delays their speech to that of infants who still do not associate their production of sounds to changes in their environment. A binary voice activity signal based on a variable threshold over the framewise energy of the input signal (Figure 1) is the only output of the system. When activated, this signal will subsequently produce a reaction in the screen of the computer in the form of animated shapes and colors. Very simple feedback is given in these games, as they are oriented to infants or children with severe disabilities. This kind of games have also been pointed out by pedagogists and

educators as useful for the early stimulation of infants with severe disorders.

*Intensity games* (Rodríguez et al., 2008) allow a patient who has just learned the ability to distinguish speech production to learn to control the volume of that production. Speech intensity is calculated as the framewise energy of the input signal and is also used for the Voice Activity Detection (VAD) in Figure 1 within the voice activity games. In intensity games, an animated character flies through a left-to-right scenario (i.e. maze) and its position in the vertical axis is proportional to the intensity of the speech production. With this strategy, the user has to modulate the intensity to avoid obstacles or interact with secondary characters on screen by raising or lowering the volume of speech.

*Breathing games* use the estimated sonority value from the Figure 1 and applies a threshold over it to detect low sonority frames associated to unvoiced segment. The detection of these unvoiced speech segments produces an animation in the screen (a character blows windmills or a ball climbs up a blowpipe) resembling traditional strategies in speech therapy to train this skill.

*Tone games* follow the same approach as intensity games but they require the user to control the fundamental frequency or pitch instead of intensity, which is also required for a correct speech production. The fundamental frequency obtained in Figure 1 is used as in the example of these games given in Figure 2, where the main character (butterfly) moves up and down as the user rises or lowers the fundamental tone to make it interact with other characters, while the pitch curve is shown on the upper right corner to help the therapist.

*Vocalization games* aim to transmit to the child the correct articulation of the vowels. With that objective, a representation of the formant map is plotted with the correct standard distribution of the vowels. As the vowel map is language-dependent, vocalization games are initially oriented to the five Spanish vowels: /*a*/, /*e*/, /*i*/, /*o*/ and /*u*/. In the games, formants are extracted with LPC analysis as shown on Figure 1 and the result is shown in the screen in the formant map, where the user can check how close that vowel is to the standard values. Vocal tract normalization would be further required to adapt the standard values of formants to every user in improved versions of the game.

All the games within the "PreLingua" framework do not require any previous configuration apart from the use of a microphone and their educative value, although not studied in this paper which is focused in ASR and PV technologies, relies on the robustness of the speech processing shown in Figure 1 and in the use of simple interfaces to provide of reinforcement and stimulation to the users (very young children with severe disabilities).


Fig. 2. Tone game in "PreLingua"

### *3.2 "Vocaliza"*

"Vocaliza" is an application oriented to the speech training of the articulation abilities of the patient in isolated words and short sentences. While focusing mostly on the articulatory side of the language like explained in Section 2.2, it also introduces the user to the semantics and syntax levels of language with different activities (Vaquero et al., 2008).

"Vocaliza" follows and operation diagram as shown in Figure 3: On the upper level, the configuration interface is the way in which the therapist creates the profiles for the different users of the application. These profiles contain all the information regarding the work of every patient with "Vocaliza" (words to practice, acoustic information and interface requirements of every child). Once a user profile is created, the core of the application are the four activities developed for speech and language therapy and the speech technologies embedded in the application to provide the user with a correct feedback. Below this structure, the user interface just requires the speech input from the patient; while the output of the system (text, audio and images) will appear automatically as the patient completes the activities, not requiring any supervision by the therapist. The activities for speech and language therapy, the use of speech technologies and the user interface in "Vocaliza" are explained

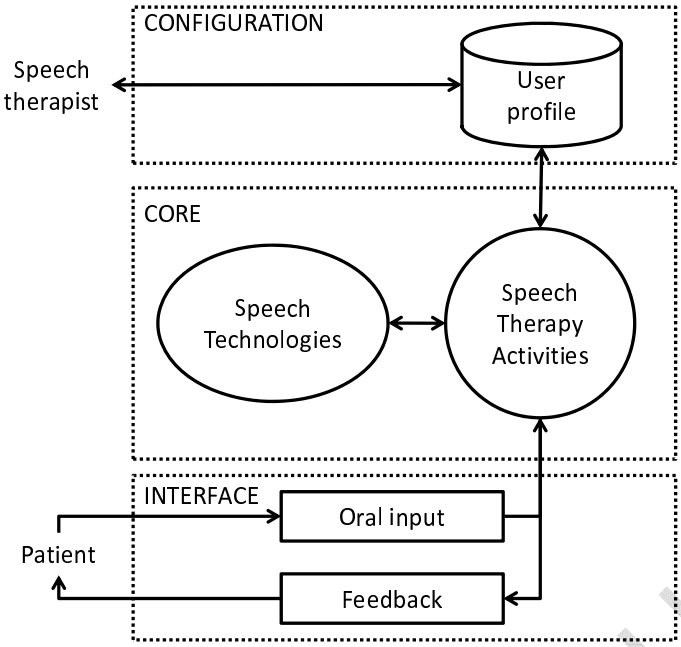


Fig. 3. "Vocaliza" operation diagram

in the following sections:

### *3.2.1 Activities for language training*

To make speech and language training attractive for young children, "Vocaliza" exercises three levels of the language (phonological, semantic and syntactic) presenting different activities.

The *phonological level* is exercised encouraging the user to utter a set of words previously selected by the speech therapist or educator during the configuration procedure to focus on the special needs of every user. The application runs an ASR decoding on the utterance to accept or reject it and evaluates the accepted utterances via a word-level PV algorithm displaying a grade as the final outcome of the game.

The *semantic level* is exercised presenting a riddle game previously defined by a speech therapist or educator. The application asks a question to the user and provides three possible answers. The user must utter the correct answer and the ASR system must accept it to go on with the next riddle. The application will show again a grade depending on the ability of the user to solve the riddle.

The *syntactic level* is exercised encouraging the user to utter a set of sentences,


Fig. 4. "Vocaliza" interface

previously selected by a speech therapist or educator. Again, the application will use ASR decoding to accept the input utterance and in that case will evaluate the utterance to display a grade.

### *3.2.2 Speech technologies for speech and language therapy*

The speech technologies provided by the core of "Vocaliza" are ASR, speech synthesis, acoustic user adaptation and PV.

*ASR* constitutes the main technology of the application. Speech therapy activities require ASR to decode user utterances, and to decide which word sequence has been pronounced so that the application will be able to let the user know if the game has been completed successfully. A robust performance of the ASR system embedded in the application is then, strongly required. This evaluation is done on Section 5 over a corpus with several impaired young speakers that will be presented on Section 4.

*Speech synthesis* provides a way to show the user how a word or sentence should be pronounced, reinforcing the correct pronunciation in the speech therapy activities. Every word, sentence and riddle is synthesized to be shown to the end user of the application during the games. The speech synthesis used within the activities is the Spanish voice of the Lernout & Hauspie Text-to-Speech (TTS) software (Coorman et al., 2000)

*Speaker adaptation* enables the application to estimate speaker-dependent acoustic models adapted to each user. "Vocaliza" uses Maximum A Posteriori (MAP) (Gauvain and Lee, 1994) estimation. Speaker adaptation is strongly

required for obtaining the full performance of the application since impaired speech can reduce dramatically the performance of ASR as it will be shown in Section 5, so that users suffering severe speech impairments would not be able to obtain any result with the application.

*PV* is the way in which the application provides an evaluation in the improvement of user communication skills. "Vocaliza" uses a word-level Likelihood Ratio (LR)-based Utterance Verification (UV)-procedure (Lleida and Rose, 2000) to assign a measure of confidence to each hypothesized word in an utterance. This method obtains the distance (as a ratio) between the likelihood of the input utterance to two models (one generated from non-impaired speech and one adapted to impaired speech). Further improved methods of phoneme-level PV are evaluated in this article in Section 6 for future implementation within the "Vocaliza" application and in future applications requiring a PV method.

### *3.2.3 User interface and AAC systems*

AAC systems and an easy user interface are the ways in which "Vocaliza" motivates children to enjoy the application while practicing their speech. The use of text, images and sounds (as in Figure 4) reinforces in the child the concept and correct pronunciation of the word or sentence that is presented.

The use of these three interactive ways of presenting every activity and the outcome of it are aimed at achieving a total accessibility by every user: Image and sound reinforce the games interface for users with low reading skills due to their developmental disabilities, while image and text provide feedback to children who may suffer hearing impairments. Finally, the use of sound and big font text and images will allow children with visual impairments to take full advantage of the potential of the application. The therapists can, hence, configure multimodality in the application to deal with the personal situation of each patient.

Furthermore, configuration of the application for the use with different users is made in a simple way to help speech therapists to work with different patients. Every patient can have a different user profile that stores all the information required, and this user profile stores how the therapist wants the words or sentences prompted to the user. Image is the prompt by default and audio and text are eligible by the therapist for every user. It also stores all the words, riddles and sentences that each user will have to practice from the whole set of words, riddles and sentences stored in the application. It also allows the therapist to decide if the ASR system embedded in the application will use speaker-independent or speaker-dependent acoustic models.

### *3.3 "Cuéntame"*

"Cuéntame" ("Tell me" in Spanish) aims at helping children with delays in the acquisition of the oral language to improve their communicative skills and follows the same philosophy that "Vocaliza" in its operation procedure as shown in Figure 3. Hence, this application also relies strongly in a robust but simple user management and configuration and in the need of using AAC systems to reinforce the correct use of language in the patients, as well as in the performance of speech technologies in speech and language therapy. As "Vocaliza", "Cuéntame" is intended to allow children to interact with the application in an unsupervised way after a short time of configuration within the application by the speech therapist.

#### *3.3.1 Activities for language training*

Three activities are designed into the application. All of them share the same vision on the initial approach that consists in scenarios of growing difficulty that the user has to solve via speech. The production of fully structured sentences is encouraged to the user in all the activities with different audio-visual rewards.

The *question answering* activities present an open ended question and the user has to provide an answer that fits the set of possible correct answers that the program has generated. The way in which the application chooses all the possible answers is shown on Figure 5: The speech therapist introduces the question that will be shown to the patient and a one-word answer to it (as the therapist only has to type one word, it simplifies the work of configuring all the activities). Then the syntax and semantic analysis over the information given by these data will generate a certain number of correct sentences to be used as possible answers. When the user gives the answer, an ASR system searchs for the keywords generated in the configuration phase.

The *descriptive* activities ask the user to describe an object accordingly to a given group of attributes (shape, color, etc...); the user has to utter a description of the object until filling up all the attributes. Once again, the user is told to use natural language and a group of possible correct sentences to describe each attribute is generated as in Figure 5.

The *dialog* activities are designed to resemble an oral command control interface as shown in Figure 6 in which a certain environment is shown to the user (house, school, shop). The user has several actions that can be done (open, take, push...) and several objects to be used (door, chair, TV...) and is asked to utter pairs of them (action-object) following a sequence of actions that lead to the achievement of the goal proposed by the application and the therapist

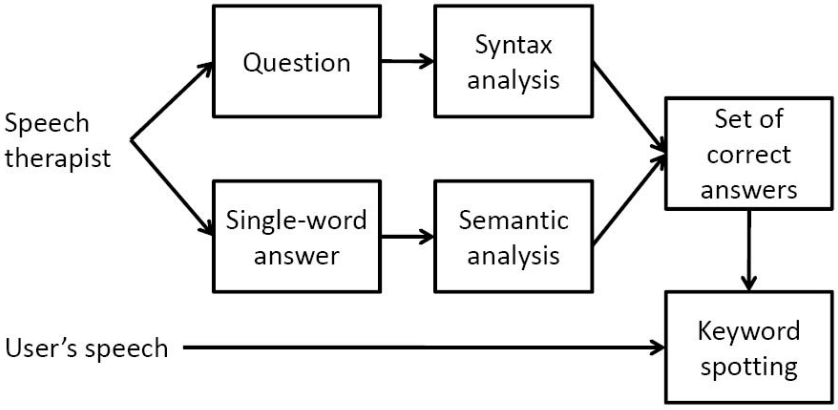


Fig. 5. Generation of possible answers in "Cuéntame"

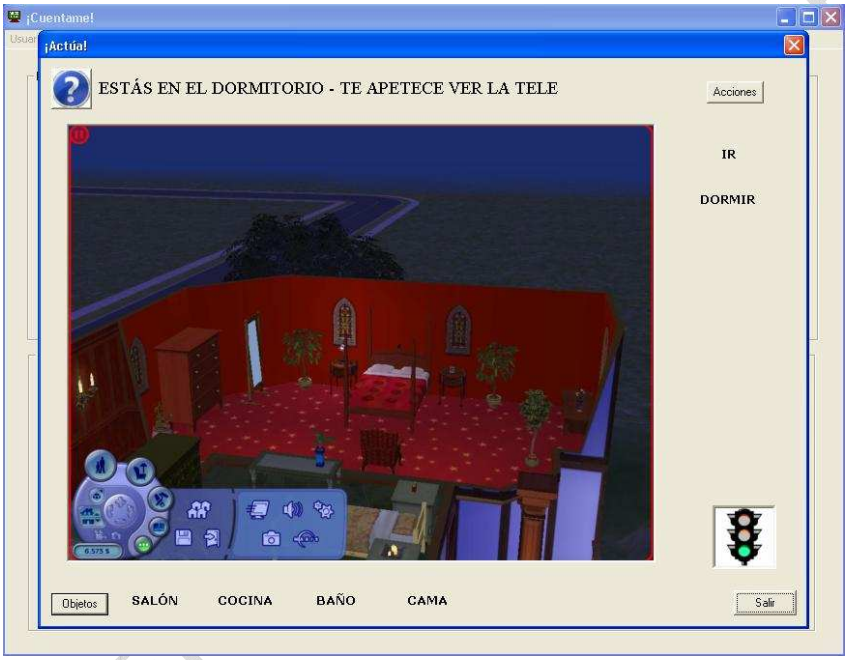


Fig. 6. "Cuéntame" interface

(for instance, turn on the TV).

### *3.3.2 Speech technologies for speech and language therapy*

"Cuéntame" shares with "Vocaliza" a similar use of speech technologies regarding ASR, PV, speech synthesis and speaker adaptation. But, further technologies are required to deal with two new issues in "Cuéntame":

First, the control and rejection of Out-Of-Vocabulary (OOV) words is required as the user is encouraged to utter freely any sentence. Hence, rejection of OOV words is necessary via a word-level confidence measure. In this case, a word-level version of the PV algorithm presented in Section 6 is used. Further-

more, correct language modeling is also required to avoid ASR errors during the activities. The language model obtained in the activities is generated directly from all the information introduced in the configuration process doing a syntactic analysis to detect the semantic categories of every word and create the language model (Figure 5). With all these keywords obtained from the morpho-syntactic analysis, the ASR system runs as a keyword spotter looking for them in the user utterance and evaluating the completeness of the sentence.

## 4 Speech Corpus of an Automated Interactive Speech Therapy Domain

In the previous Section, several CASLT applications have been presented to provide an easy way of unsupervised speech and language therapy. Hence, the need for configuration of the speech technologies (ASR and PV) used in the speech therapy domain arises. Although all the applications in Section 3 are aimed to the population of children with different degrees of neuromuscular disorders and language impairments, the focus in this Section will be set in the characterization of the "Vocaliza" domain in Section 3.2. "Vocaliza" was the firstly developed tool in "Comunica" and has been the object of the research shown in this paper during the past years. This tool asks the child to utter a set of words chosen by the therapist and evaluates them based on the output of an ASR decoding and PV algorithm. This characterization of the domain is done and explained in Section 4.1 via the acquisition of a novel corpus containing impaired speech from children and young adults in the form of isolated words uttered during the use of the "Vocaliza" application.

Furthermore, it is strongly required to evaluate an efficient mechanism to provide the evaluation of impaired speech. This can only be obtained via a manual labeling of the collected speech that is explained in Section 4.2, altogether with the acquisition of a corpus with unimpaired speech from young children in Section 4.3 for the task and domain adaptation of the systems.

### *4.1 A novel corpus of impaired speech from children and young adults*

The acquisition of this novel corpus (Saz et al., 2008) had to follow the same process that the users have on the "Vocaliza" tool and had to be realistic in the speech impairments that actual users of "Vocaliza" might have. Regarding disordered speech, some previous corpora have been acquired for studies in dysarthic speech like Nemours (Menéndez-Pidal et al., 1996), a research corpus of 814 short sentences in English from 11 dysarthric speakers, and the corpus for the STARDUST project (Hawley et al., 2003) used for the

| Word | SAMPA | Word | SAMPA | Word | SAMPA |
|---|---|---|---|---|---|
| árbol | /arBol/ | boca | /Boka/ | bruja | /Bruxa/ |
| cabra | /kaBra/ | campana | /kampana/ | caramelo | /karamelo/ |
| casa | /kasa/ | clavo | /klaBo/ | cuchara | /kutSara/ |
| dedo | /DeDo/ | ducha | /DutSa/ | escoba | /eskoBa/ |
| flan | /flan/ | fresa | /fresa/ | fuma | /fuma/ |
| gafas | /Gafas/ | globo | /Globo/ | gorro | /Gorro/ |
| grifo | /Grifo/ | indio | /indjo/ | jarra | /xarra/ |
| jaula | /xawla/ | lápiz | /lapiT/ | lavadora | /laBaDora/ |
| luna | /luna/ | llave | /LaBe/ | mariposa | /mariposa/ |
| moto | /moto/ | niño | /niJo/ | ojo | /oxo/ |
| pala | /pala/ | palmera | /palmera/ | pan | /pan/ |
| peine | /pejne/ | periódico | /perjodiko/ | pez | /peT/ |
| piano | /pjano/ | pie | /pje/ | piña | /piJa/ |
| pistola | /pistola/ | plátano | /platano/ | playa | /plaLa/ |
| preso | /preso/ | pueblo | /pueBlo/ | puerta | /pwerta/ |
| ratón | /raton/ | semáforo | /semaforo/ | silla | /siLa/ |
| sol | /sol/ | tambor | /tamBor/ | taza | /taTa/ |
| teléfono | /telefono/ | toalla | /toaLa/ | toro | /toro/ |
| tortuga | /tortuga/ | tren | /tren/ | zapato | /Tapato/ |

Table 1
Words in the Induced Phonological Register and their SAMPA transcription

development of speech therapy tools and augmented communication devices for heavily impaired individuals. The Universal Access Database (Kim et al., 2008) is a massive corpus containing several hours of speech from 10 individuals with different degrees of dysarthria. In languages different from English, it can be referred also efforts of speech acquistion in Dutch (Sanders et al., 2002), and regarding Spanish, the most remarkable effort was done in the HACRO project (Navarro-Mesa et al., 2005), containing speech from speakers of different ages with different speech impairments for the development of a tool for the evaluation of oral utterances. This corpus was the model for the acquisition of the corpus used in this work, although it is expanded in the fact that several sessions are recorded from every speaker and the disabilities of the speakers are mostly focused on developmental delays and the speech impairments associated to them.

The procedure in the acquisition of every session was to place every speaker in an scenario with the "Vocaliza" application in which a certain set of words of special speech therapy interest was presented. The application was running on adaptation mode, a mode that allows the user to adapt the acoustic models to their own speech by recording several utterances. In this mode, no ASR is run and every utterance can be visually checked in its waveform to evaluate the quality of the recorded signal. The set of words used for the recordings was the Induced Phonological Register (RFI) (Monfort and Juárez-Sánchez, 1989). RFI, while containing only 57 words is a powerful set of words for speech therapy as it contains examples of all the 23 phonemes and 70% (36 out of 51) of the allophones described traditionally in the Spanish language (Alarcos, 1950) as shown in Table 1. The total amount of syllables in the 57 words is 129 (an average of 2.26 syllables per word, with 90 different syllables) and the number of phonemes is 292 (an average of 5.13 phonemes per word).

From the technical aspect, a wireless close-talk microphone was used in the recording. This way, comfortability of the speakers was guaranteed as they were not directly attached to the computer while obtaining the best speech quality possible with a high Signal-to-Noise Ratio (SNR). More precisely, the recordings in the impaired speech corpus have an average SNR of 26.4 dB. The recordings were made in an empty classroom environment in the "Alborada" school while the rest of the classes were running normally in the school. The speaker was told to repeat the utterance if an excess of noise from the environment was captured.

The number of speakers that participated in the acquisition was 14; 7 males and 7 females distributed in ages from 11 to 21 years as shown in Table 2. All of these 14 speakers recorded 4 sessions of the isolated words to make a total of 3 192 isolated word utterances (2 hours and 56 seconds of speech including silence). Every session was acquired in different days to reflect intra-speaker variability. All the speakers have been diagnosed by their educators and speech therapists with different levels of dysarthria due to several origins as well as other language disorders at the semantic/syntactic and pragmatic levels.

### *4.2 Manual evaluation of the impaired speech*

Once the corpus was acquired, the evaluation of any robust ASR system in Section 5 or PV algorithm in Section 6 required of a manual labeling in the corpus that validated the systems. Thus, a labeling process was started. In this process, every phoneme in the corpus was labeled by three different transcribers as having been either deleted, mispronounced and therefore substituted with another phoneme, or correctly pronounced. In the end, the final label for the phoneme was chosen by consensus among the labelers. The average percentage

Impaired Speech Corpus

| Code | Gender | Age | Code | Gender | Age | Code | Gender | Age |
|---|---|---|---|---|---|---|---|---|
| Spk01 | F | 13 | Spk06 | M | 16 | Spk11 | F | 19 |
| Spk02 | M | 11 | Spk07 | M | 18 | Spk12 | M | 18 |
| Spk03 | M | 21 | Spk08 | M | 19 | Spk13 | F | 13 |
| Spk04 | F | 20 | Spk09 | F | 11 | Spk14 | F | 11 |
| Spk05 | M | 18 | Spk10 | F | 14 | - | - | - |

Table 2
Speakers in the corpus: Code, gender (Male or Female) and age

of correct phonemes is 82.4%, while 10.3% of the phonemes are substituted and 7.3% are deleted and the results for every speaker are shown in Table 3. The total number of labelers was 10, all of them with expertise in the fields of speech technologies or phonetics.

The labeling strategy resembles more lexical labeling than speech quality labeling. This was originally intended to create a more objective measure of the mispronunciations by the speakers in the corpus. Also, with this type of labeling, labels are more consistent as the pair-wise inter-labeler agreement rate is 85.8% which raises to 89.7% when considering only a binary decision: Correct versus Incorrect (deletions plus substitutions). This consistent labeling avoids the problems of a subjective speech quality measurement that would have required very experienced labelers and would have suffered more of subjective differences within the evaluation given by different labelers.

This lexical labeling also aims to assimilate the approach to the speech disorder that a speech therapist might have when dealing with impaired speech from young disabled patients. In this situation, the transmission of the message is more important than speech quality, and this labeling procedure focuses on the substitution of phonemes that would cause a change in the meaning of the uttered word. Incorporating a PV algorithm that correctly detects this kind of mispronunciation in tools like "Vocaliza" and "Cuéntame" would help speech therapists in their daily work with this kind of disorders. The effort in this task will show its results in Section 6.

### *4.3 An unimpaired corpus for domain adaptation*

As a parallel process to the recordings of the impaired speech corpus, recordings of a reference corpus containing speech from unimpaired speakers in the same age range as the impaired speakers were made. This reference corpus was considered necessary to avoid the mismatch due to the difference of age

Human Labeling

| Speaker | Del. | Subs. | Corr. | Speaker | Del. | Subs. | Corr. |
|---|---|---|---|---|---|---|---|
| Spk01 | 0.2% | 0.9% | 98.9% | Spk08 | 13.1% | 17.7% | 69.2% |
| Spk02 | 9.2% | 12.4% | 78.4% | Spk09 | 2.9% | 5.3% | 91.8% |
| Spk03 | 0.7% | 4.6% | 94.8% | Spk10 | 8.4% | 13.1% | 78.5% |
| Spk04 | 1.1% | 2.1% | 96.8% | Spk11 | 2.1% | 4.7% | 93.2% |
| Spk05 | 17.4% | 26.1% | 56.5% | Spk12 | 11.7% | 14.0% | 74.3% |
| Spk06 | 0.2% | 0.5% | 99.3% | Spk13 | 25.9% | 30.5% | 43.6% |
| Spk07 | 5.6% | 7.3% | 87.1% | Spk14 | 3.9% | 5.1% | 91.0% |
| AVG. | 7.3% | 10.3% | 82.4% | | | | |

Table 3
Labeling results per speaker: Rate of deletions, substitutions and correct phonemes

between the impaired speakers and the age of the speakers contained in the adult corpora used for creating the speaker-independent models for the experiments in ASR in Section 5.1 and in PV in Section 6.2. This mismatch due to age could mask the mismatch in speech due to the impairments of the speakers that gets better reflected when comparing speech from subjects in the same age range, providing a very useful task and domain adaptation as shown in Section 5.2.

This unimpaired corpus should repeat the same acquisition scenario as the impaired speech corpus. Thus, the same vocabulary (RFI) and the same type of sessions (isolated words) were chosen for this recording scenario. The amount of speakers in this corpus is 168, 73 boys and 95 girls ranging in the age from 10 to 18 years. Every speaker uttered a session of the 57 words in the RFI, which makes a total number of 9 576 isolated-word utterances in the corpus (6 hours, 17 minutes and 43 seconds of speech including silence). The recording process was exactly the same, using "Vocaliza" as the domain of acquisition and with a close-talk microphone that guaranteed a high average SNR of 25.6 dB.

## 5 Experiments on Dysarthric ASR

ASR is the core of the "Vocaliza" and "Cuéntame" applications. A robust performance of the ASR system is not only important for CASLT tools, but all the knowledge gained in these tests can be also used for future works in oral command control systems for physically handicapped individuals (Hawley et al., 2003). These systems require of ASR to improve the performance in a

task with the type of speech presented in Section 4, as the speakers might also be tentative users of this kind of technical aids due to their physical disabilities. The results shown in this Section present a baseline of ASR that could be used within this line of work.

Three different experiments were designed: First, results with a speaker-independent ASR system trained from a population of adult speakers were obtained. Subsequently, a task-dependent but speaker-independent ASR system was tested. And finally, several speaker-dependent ASR systems for every impaired speaker were evaluated. This line of work (from speaker independence to speaker dependence via task domain adaptation) will be also be followed by the PV algorithms in Section 6. All these experiments were required to understand how ASR and PV react to impaired speech from children and young adults with different impairments. These speakers are users of applications like "Vocaliza" in Section 3.2 and "Cuéntame" in Section 3.3 and their speech was collected as explained in Section 4.

### *5.1 Speaker-independent results*

First, all the isolated-word utterances in the corpus were evaluated through a speaker-independent ASR system. This model was trained with the Spanish unimpaired adult speech corpora Speech-Dat Car (Moreno et al., 2000), Albayzín (Moreno et al., 1993) and Domolab (Justo et al., 2008) via Maximum Likelihood (Dempster et al., 1977) optimization. The model set, and the ones generated from it, consisted of a set of 746 context-dependent units plus a begin-end silence model; every model being a 1-state Hidden Markov Model (HMM) whose distribution is a Gaussian Mixture Model (GMM) made up of 16 Gaussians. Features for the ASR system were extracted every 10 milliseconds using a 25-millisecond Hamming window; 12 Mel-Frequency-Cepstrum Coefficients (MFCC) were then obtained and were used within the ASR system plus their first and second time derivatives; log-energy plus its first and second derivatives are also calculated to create a final 39-dimensional feature vector.

Results for the speaker-independent set of experiments are shown in Table 4 for every one of the 14 impaired speakers shown in Section 4.1. The average Word Accuracy (WAC) for all the speakers reached 66.8%, 29.94% lower than the results obtained using the same ASR system for the 168 unimpaired speakers, whose average WAC was 96.70%. These results show the big influence of the disorders suffered by the speakers, specially for the most impaired speakers like $Spk05$, $Spk12$ or $Spk13$ whose WAC was below 50%, according to the results in the labeling in Section 4.2 that showed these speakers as some of the ones with more problems in their speech production and intelligibility.

| | Speaker-Independent ASR | | | | | | |
|---|---|---|---|---|---|---|---|
| | Spk01 | Spk02 | Spk03 | Spk04 | Spk05 | Spk06 | Spk07 |
| WAC | 99.6% | 64.5% | 72.8% | 85.1% | 38.2% | 93.9% | 73.3% |
| | Spk08 | Spk09 | Spk10 | Spk11 | Spk12 | Spk13 | Spk14 |
| WAC | 56.6% | 71.9% | 61.0% | 89.0% | 30.7% | 21.5% | 76.8% |
| Avg. | 66.8% | | | | | | |

Table 4
Speaker-independent ASR results

| | Task-Dependent ASR | | | | | | |
|---|---|---|---|---|---|---|---|
| | Spk01 | Spk02 | Spk03 | Spk04 | Spk05 | Spk06 | Spk07 |
| WAC | 92.9% | 84.2% | 92.5% | 98.3% | 43.4% | 96.9% | 76.3% |
| | Spk08 | Spk09 | Spk10 | Spk11 | Spk12 | Spk13 | Spk14 |
| WAC | 55.3% | 77.6% | 71.1% | 94.3% | 37.7% | 23.7% | 82.5% |
| Avg. | 73.6% | | | | | | |

Table 5
Task-dependent ASR results

### *5.2 Task-dependent results*

The task and domain adaptation was achieved adapting the initial speaker-independent model trained with adult speech using the unimpaired speech corpus explained in Section 4.3 via a Maximum A Posteriori (MAP) strategy (Gauvain and Lee, 1994). The MAP algorithm guaranteed a good convergence with an amount of data like the one available for the task domain adaptation, more than 6 hours of speech, which was entirely used for the adaptation.

The results are shown in Table 5 and gave an average WAC of 73.6%. The increase of WAC for the task and domain adapted ASR is a significant 22.92% but it is still far from the 96.70% WAC achieved by the unimpaired speakers. This relative task-domain adaptation increase ($TDINC$) was calculated as follows:

$$TDINC(\%) = \frac{TDWAC - SIWAC}{UWAC - SIWAC} 100\% \quad (1)$$

where $TDWAC$ is the WAC with the task-dependent acoustic model, $SIWAC$ is the WAC with the initial speaker-independent acoustic model and $UWAC$ is the reference WAC of the unimpaired speakers.

The difference in WAC between the speaker-independent model (66.8%) and

| | Speaker-Dependent ASR | | | | | | |
|---|---|---|---|---|---|---|---|
| | Spk01 | Spk02 | Spk03 | Spk04 | Spk05 | Spk06 | Spk07 |
| WAC | 99.6% | 90.8% | 99.6% | 99.6% | 54.8% | 100.0% | 93.9% |
| | Spk08 | Spk09 | Spk10 | Spk11 | Spk12 | Spk13 | Spk14 |
| WAC | 76.3% | 90.4% | 89.0% | 99.1% | 82.9% | 30.3% | 96.0% |
| Avg. | 85.9% | | | | | | |

Table 6
Speaker-dependent ASR results

the task-dependent model (73.6%) relied in the mismatch between the adult speech and the children and young speakers' speech used for training both models, as well as in the adaptation to the task (57 words in the RFI) and to the acoustic conditions of the recordings. The 73.6% WAC was only affected in this case by the speech disorders of the speakers and was shown to be significantly higher than the WAC obtained with unimpaired speakers.

### *5.3 Speaker-dependent results*

Speaker-dependent ASR experiments were carried out within a leave-one-out strategy in which three of the isolated-word sessions of every speaker were used for adaptation and the remaining session was used for evaluation. Finally, the definitive WAC for every speaker was obtained as the average WAC for the 4 evaluated sessions. The algorithm used for adaptation was again a MAP approach (Gauvain and Lee, 1994) using as seed the task-dependent model trained for the experiments in Section 5.2. In this case there are no units appearing in the test data that are unseen in the adaptation data, so it was considered not necessary the use of Maximum Likelihood Linear Regression (MLLR) (Legetter and Woodland, 1995) to characterize the speaker-dependent models.

Results for the speaker-dependent set of experiments are shown in Table 6 for all the impaired speakers. The average WAC increased to 85.9% (63.83% of relative improvement over the speaker-independent results) This relative speaker adaptation increase ($SDINC$) was calculated as follows:

$$SDINC(\%) = \frac{SDWAC - SIWAC}{UWAC - SIWAC} 100\% \quad (2)$$

where $SDWAC$ is the WAC with speakers-dependent acoustic models, $SIWAC$ is the WAC with the initial speaker-independent acoustic model and $UWAC$ is the reference WAC of the unimpaired speakers.

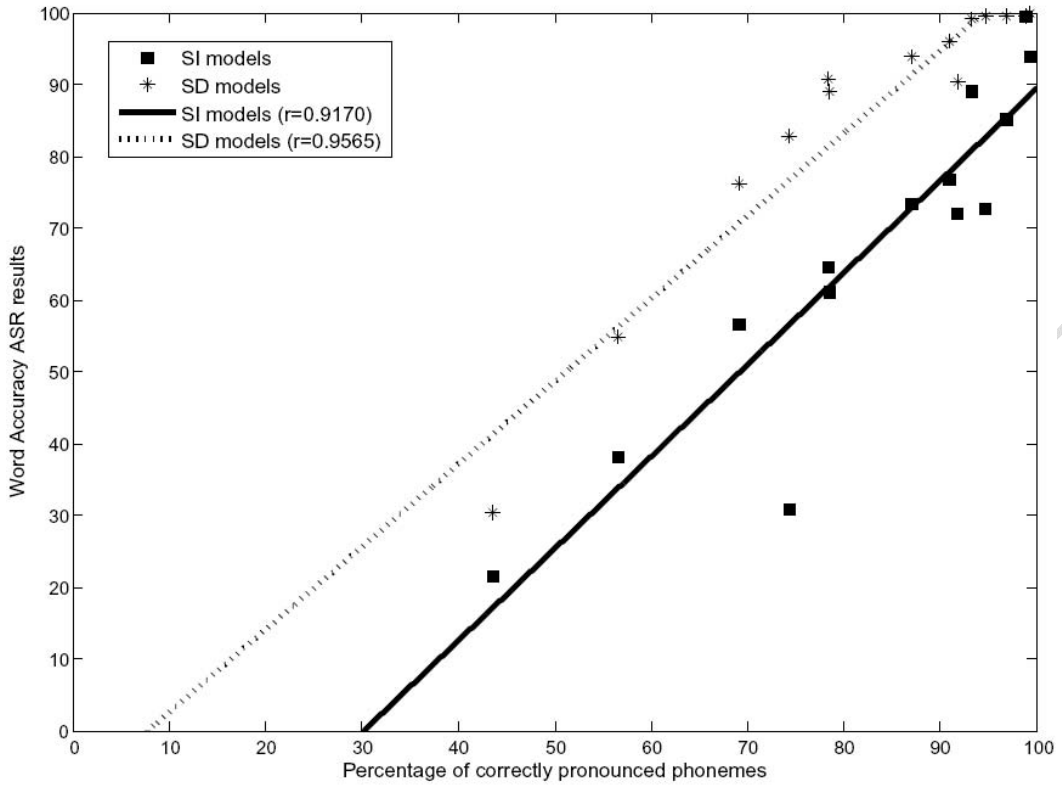


Fig. 7. Scatter plot correlating the percentage of correct phonemes for every speaker with the WAC results in the SI and SD cases, and the linear regression estimates

In this case, the increase was more significant than in the task-dependent situation, and was rising towards the WAC obtained by the unimpaired speakers (96.70%), but still a 10.83% lower. It was to notice that half of the speakers had a WAC around or below 90% even with speaker adaptation.

### *5.4 ASR for Speech and Language therapy*

One of the design goals of the speech therapy tools was that they were mainly intended to be used by patients with cognitive disorders (as most of the pupils in the CPEE Alborada). This situation leaded to the fact that the feedback in the pronunciation games should be given in a simple way, so it could be easily understood by any user. The idea in "Vocaliza" was that ASR could be a simple and robust way of pronunciation verification; as it would reject utterances which were extremely distorted from the canonical pronunciation.

This section aims to test if ASR is a good word-level PV method to reject heavily mispronounced utterances. The proposed scenario is as follows: The speech therapists configures the tool with a certain number of words to be trained by a given patient. Each word is presented isolated to the patient

who has to utter it as correctly as possible and the ASR system embedded in the tools decodes the most likely pronounced word from the vocabulary inserted by the therapist. If the output of the ASR system is the same that the word presented to the patient, it proceeds to the next word (accepting the utterance of the user). If the ASR output is different, the word is considered mispronounced (rejecting the utterance of the user) and prompted again.

Main goals of the tests were to obtain a low rejection of correctly pronounced utterances (to avoid frustration in the user) and a low acceptance of incorrect utterances (to force the user to improve the pronunciation). This possibility arises watching at the correlation between the correctness of the speakers' pronunciation (seen in Table 3) and the ASR results (Tables 4, 5 and 6). The scatter plot between these two measures is shown in Figure 7 for the speaker-independent and speaker-dependent cases, where the linear regression functions are also plotted. The correlation coefficients ($r$) are 0.917 for the speaker-independent regression model and 0.9565 for the speaker-dependent model, which shows the correlation of ASR results with the labeling results.

Hence, two new measures of ASR performance are used, being these the False Rejection Ratio (FRR) in Equation 3 and the False Acceptance Ratio (FAR) in Equation 4:

$$FRR = \frac{RCU}{CU} \tag{3}$$

$$FAR = \frac{AIU}{IU} \tag{4}$$

with $CU$ being the number of correct utterances and $IU$ the number of incorrect utterances. $RCU$ is the number of correct utterances that have been rejected and $AIU$ the number of incorrect utterances that have been accepted.

One key point for this measure was to determine what could be considered a mispronounced utterance. Every utterance contained a single isolated word; so, as utterances were labeled at the phoneme level but not at the word level, a mapping from phoneme labels to word labels was required. Initially, every utterance with any number of mispronounced phonemes was considered as mispronounced; this is, only words correct in all their phonemes were considered as correct words.

These results are presented in the first columns of Table 7 (Case 1). With the use of speaker-independent acoustic models, the FRR was 12.3% and the FAR was 43.3%. FRR can be reduced to 1.5% with the use of speaker-dependent acoustic models as seen on Section 5.3, but the FAR rises to 70.5%. This value of FAR could not be considered a reliable operation point for any specialist in speech therapy, as it was losing all of its pedagogical interest. Anyway,

| | Case 1 | | | Case 2 | | |
|---|---|---|---|---|---|---|
| | SI model | TD model | SD model | SI model | TD model | SD model |
| FRR | 12.3% | 5.5% | 1.5% | 29.4% | 15.5% | 5.8% |
| FAR | 43.3% | 53.6% | 70.5% | 27.1% | 32.4% | 52.3% |

Table 7
FRR and FAR of speech recognition as pronunciation verification method

it was to be noticed that those words with any number of mispronounced phonemes were considered as mispronounced words. Considering that the RFI vocabulary was relatively small (57 words in Table 1) there were no minimal pairs between them (this is, words with just one phoneme different); so, it was expectable that words with only one mispronounced phoneme were correctly recognized by the system.

To study this influence, a second experiment (Case 2) was performed. In this case, a word was only considered as mispronounced when a minimum of two if its phonemes were labeled as mispronounced. Results in terms of FRR and FAR are given in the last columns of Table 7. In this case, an Equal Error Rate (ERR) situation was achieved when using speaker-independent acoustic models (FRR of 29.4% and FAR of 27.1%). This operating point was acceptable, but possible frustration could still be lowered by the use of speaker-dependent acoustic models. In this case, FAR rose to 52.3%, so there was a big number of incorrectly pronounced utterances who were being accepted by the system.

The main conclusion that can be achieved from these experiments is that the presented word-level PV is not an accurate situation for robust speech therapy (although considered necessary by the therapists when dealing with individuals with cognitive disorders). Hence, the need for phoneme-level PV arises, being studied in Section 6 as a work necessary in the development of tools oriented to provide a feedback with more precise resolution.

## 6 Verifying Pronunciations Arising from Dysarthric Speech

This section describes a set of confidence measure-based techniques for detecting phoneme-level mispronunciations in utterances from the impaired population described in Section 4. The task domain is defined by the user interface in the current configuration of the interactive learning tool "Vocaliza", which was introduced in Section 3.2. In this task domain, all utterances correspond to isolated words taken from the 57 word phonetically balanced RFI vocabulary (Monfort and Juárez-Sánchez, 1989). User utterances are elicited in the interactive dialog using written and pictorial prompts, as described in Section 3. As a result, PV can be assumed to be a problem of verifying the claim

that a particular phoneme occurring in an utterance of a known word has been correctly pronounced.

This PV scenario is similar to the problem of PV for automated language learning and language skills evaluation applications (Zhang et al., 2008). However, members of the young impaired speaker population suffer from neuromuscular disorders of varying severity. This distinguishes this population from language learners who are assumed to not be proficient in the given language but at the same time are assumed to not suffer from any speaking impairments. Hence, speech obtained from the impaired population of speakers is more likely to be significantly affected at multiple levels than speech from unimpaired speakers. Effects of these disorders can be observed in frame level spectral characteristics, segment level coarticulation, lexical level pronunciation rules, and suprasegmental pitch contours (Patel, 2002). While there has been considerable effort made to model how these disorders are reflected in the underlying articulatory dynamics of speech production (Deller et al., 1991), the techniques described here are based on a posteriori probabilities derived from HMM-based ASR. Phoneme-level measures of confidence are derived from the acoustic speech utterance and are used to define a decision rule for accepting or rejecting the hypothesis that a phoneme was mispronounced.

There are three important issues that are addressed in this section (Yin et al., 2009). First, a phoneme-level confidence measure is derived that incorporates both acoustic and non-acoustic knowledge sources that may be impacted by the mispronunciations resulting from the speech impairments. In Section 6.1, a procedure used for deriving confidence measures based on posterior probabilities derived from phoneme lattices is described (Mangu et al., 2007). Second, in order for a confidence measure to be effective in detecting variability arising from the mispronunciations produced by the impaired speaker population, it is necessary to reduce the influence that other sources of variability have on the confidence measure. In Section 6.2, the application of acoustic HMM adaptation to limit the effects of interspeaker variability and task variability is described.

The last issue concerns the fact that the PV task predicts the potentially subjective decisions made by human labelers working under the labeling scheme given in Section 4.2. In Section 6.3, a nonlinear mapping is performed to translate the a posterior probabilities estimated from ASR lattices to confidence measures that can better predict these manually derived labels.

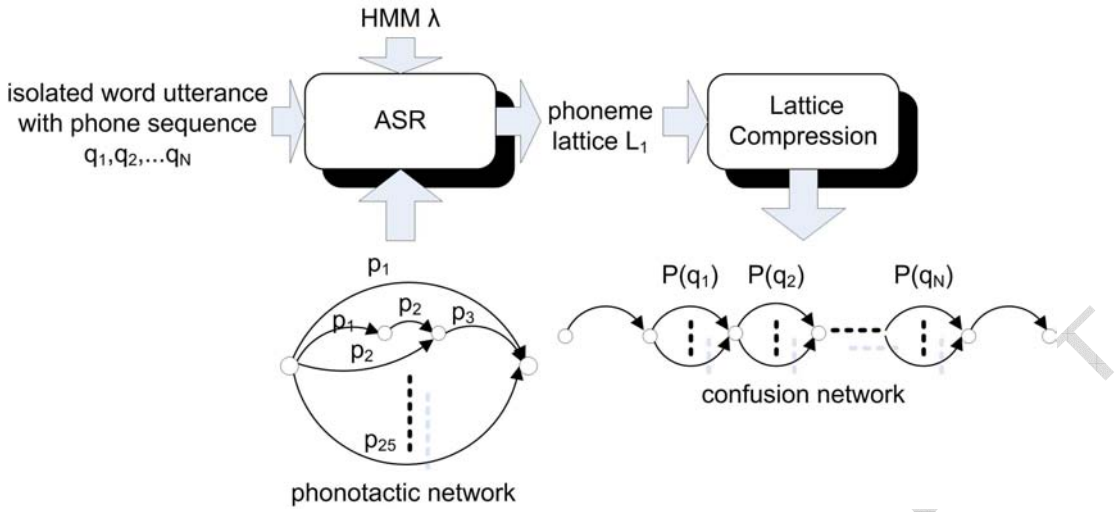


Fig. 8. Confusion network-based posterior probability estimation

### *6.1 Phoneme Level Confidence Measure*

In the phoneme PV scenario, it was assumed that the "target" word sequence and its baseform phonetic expansion were known. For the experimental study described in Section 6.4, it was assumed that the input test utterance corresponded to an isolated word and the corresponding baseform phoneme string, $q_n$, $n = 1, \ldots, N$, was known. PV in this context simply refers to obtaining confidence measures for each phoneme in the baseform expansion and applying a decision rule for accepting or rejecting the hypothesis that a given phoneme was correctly pronounced. The process, as depicted in Figure 8, was performed in two steps.

First, phoneme recognition was performed on the given isolated word utterance where search was constrained using a network that described the potential pronunciations that might be expected from an unimpaired speaker. This network could potentially be created from the syllabification rules of the language or be trained from observed pronunciations decoded from the population of unimpaired speakers. While rule-based constraints are currently being investigated, simple n-gram phonotactic constraints were applied here. Specifically, a bigram phonotactic model was trained from baseform phonetic expansions obtained from an 8 million word subset of the Spanish language section of the Europarl speech corpus (Koehn, 2005), containing transcriptions from several sessions in the European Parliament translated to different European languages. This phonotactic bigram model was also used to constrain search as outlined in Figure 8.

Second, a confusion network, as depicted in Figure 8, was created using a lattice compression algorithm (Mangu et al., 2007). A phoneme lattice $L_1$ which contained phoneme labels and their associated acoustic and language proba-

bilities on the arcs was generated by the ASR decoder. A phoneme lattice is a Directed Acyclic Graph (DAG) which contains a large number of competing phoneme hypotheses. These lattices generally contain a large number of redundant paths and can often be very large. The confusion network aligned links on the phoneme lattice and transformed the lattice into a linear graph in which all arcs that emanated from the same start node terminated in the same end node. The ordering properties of the original lattice were maintained in the confusion network. The posterior phoneme probabilities $P(q_n)$, $n = 1, \ldots, N$, appearing on the transitions of the confusion network were obtained by forming the sum of probabilities of all the paths passing through the target phoneme arc in the lattice and normalizing by the sum of probabilities of all the paths in the lattice. Compared with the original phoneme lattice, the compact structure and the ordering properties of the confusion network facilitated efficient evaluation of posterior-based confidence measures for verifying phoneme pronunciations. The posterior probabilities for the baseform phonemes in the target phoneme string were obtained by aligning the target phoneme string with the phoneme lattice $L_1$. These posterior phoneme probabilities were used as phoneme-dependent confidence scores. A decision criterion for verifying whether a given target phoneme had been correctly pronounced could be implemented by comparing these scores with a decision threshold.

### *6.2 Reducing variability through model adaptation*

Adaptation scenarios are presented here for reducing the effects of other sources of variability. These may include all sources of variability outside of those introduced by the speech disorders existing among the disabled speaker population. For example, physiological and dialect differences among speakers, differences in microphones, and differing acoustic environments can all influence the ability to detect mispronunciations.

The baseline ASR system and the adaptation scenarios included in our experimental study described in Section 6.4 are introduced. Baseline HMM models were trained from the Spanish language Albayzín speech corpus (Moreno et al., 1993), which includes 6 800 sentences with 63 193 words. This corpus contains 6 hours of speech including silence; however, only 700 unique sentences are contained in the corpus. Because of this lack of phonetic diversity, it was difficult to train context-dependent models that could generalize across task and domains. For this reason and because of the simplicity of this small vocabulary task, context-independent monophone models were used here. In all experiments, 25 monophone-based context-independent HMMs were used which consisted of 3 states per phoneme and 16 Gaussians per state. MFCC observation vectors along with their first and second difference coefficients were used as acoustic features. Furthermore, the corpora used for the train-

ing of speaker-independent models in Section 5.1 like SpeechDat-Car (Moreno et al., 2000) and Domolab (Justo et al., 2008) were not used at this point as they are not designed to be phonetically balanced and do not relate to this task.

Phoneme-level PV was performed on isolated word utterances from a 57 word vocabulary where each utterance was an average of only 2.3 seconds in length including silence. In order to obtain a more robust task-dependent acoustic model, the unimpaired corpus described in Section 4.3 was used to perform a MAP-based (Gauvain and Lee, 1994) and MLLR transform-based (Legetter and Woodland, 1995) adaptation of the Gaussian mean vectors. The MLLR adaptation involved two regression classes, one for the silence and the other one for all the non-silence phonemes. The reason for combining both MAP and MLLR adaptation was based on their complementary behavior (Goronzy and Kompe, 1999). MAP adaptation was performed independently on the means associated with distributions assigned to each phoneme classes. Improved models were obtained using MAP for phonemes that were well represented in the adaptation data. On the other hand, MLLR adaptation was applied as a linear transformation to the mean vectors of the distributions. MLLR has the ability to benefit from observation vectors belonging to all phoneme classes to adapt those models that are not well represented in the adaptation data. As a result, simply combining the two adaptation procedures can result in complementary performance increases.

The MAP/MLLR task-dependent adaptation corpus included 6 840 adaptation utterances spoken by 120 unimpaired speakers from the 168 children and young adults in Section 4.3. Each unimpaired speaker provided 57 isolated test word utterances where all the words were assumed to be accurately pronounced. The adaptation corpus contained 4.5 hours of speech including silence.

Supervised speaker-dependent adaptation for each test speaker was performed using an MLLR-based transform applied to the Gaussian means of the task-dependent HMM. For each speaker, a single MLLR transform matrix was estimated and applied for speaker adaptation. The speaker-dependent MLLR adaptation data consisted of 57 isolated word utterances, or 2.2 minutes of speech for each of the test speaker. The remaining 2 394 impaired speakers utterances, three sessions of 57 isolated word utterances for each impaired speaker, were used for evaluation. The supervised speaker-dependent MLLR transformation was then applied prior to verifying the phoneme level pronunciation of the impaired speech utterances.

### *6.3 Non-linear Mapping of Posterior Probabilities*

A nonlinear transformation was performed to map the lattice posterior probabilities to phoneme-level confidence measures. There were two motivations for this: The first motivation stemmed from the fact that all of the PV techniques presented here were evaluated in terms of their ability to predict the labels defined by the labeling scheme seen in Section 4.2. The decision made by an expert as to whether a given occurrence of a phoneme was classified as being "mispronounced" rather than as a "pronunciation variant" will always have a subjective component. The labeling scheme presented in Section 4.2 is important because it addresses the trade-off between the need for a consistent, repeatable, and easily implemented labeling strategy against the need for an accurate characterization of the quality of pronunciation of a given phoneme. There is no guarantee, however, that the posterior probabilities estimated as shown in Figure 8 will always be accurate predictors of these labels.

The second motivation for mapping the lattice posterior probabilities to phoneme-level confidence measures was the fact that there was a great deal of prior information available in this PV scenario. This included knowledge of the target word, the target phoneme, and the position of the phoneme within the word. This prior information could be combined with the phoneme-level posterior probability using one of many possible fusion strategies to better predict the human derived labels.

In the experimental study described in Section 6.4, the parameters of a single-layer multilayer perceptron with the above parameters as input were trained to implement a non-linear transformation. Backpropagation training was performed for a network with 47 hidden nodes and with input activations which included the phoneme-level posterior probabilities, indicator variables corresponding to each of the phoneme labels, and indicator variables corresponding to the word labels. The network was trained with the human derived pronunciation labels serving as targets. PV was performed using the output activations obtained from this network on test utterances with the same kinds of input parameters as the ones used in the training phase.

### *6.4 Study of PV Performance*

This section presents an experimental study performed to evaluate the ability of the PV techniques presented in Section 6 to detect mispronunciations in utterances obtained from impaired speakers. Verification performance is measured using utterances from the 14 speaker population of impaired speakers as a test corpus. For each phoneme in the baseform phonetic expansion of a

| Phoneme-level Verification Performance (EER) | | |
|---|---|---|
| Adaptation Scenario | zerogram | bigram |
| TIND HMM (Baseline) | 25.3% | 22.2% |
| TDEP MAP/MLLR Adaptation | 19.7% | 18.4% |
| SDEP MLLR Adaptation | 18.3% | 17.1% |
| SDEP NN Mapping | 14.9% | N/A |

Table 8
Phoneme detection performance measured as the equal error rate (EER) for task-independent (TIND) baseline, task-dependent (TDEP) MAP/MLLR adaptation, speaker-dependent (SDEP) MLLR adaptation, and SDEP NN-based non-linear mapping.

word, the task is to verify the claim that the pronunciation of that phoneme is correct according to the human labels assigned using the labeling scheme described in Section 4.2. Since this is in fact a detection problem, the performance is presented using receiver operating characteristic (ROC) curves as the probability of false rejection (FR) vs. the probability of false acceptance (FA) of a pronunciation hypothesis. These characteristics can be summarized using the equal error rate (EER) measure. The EER is computed by applying a threshold to the phoneme-level confidence scores and identifying the threshold setting where the probability of false acceptance is equal to the probability of false rejection.

The performance relating to several issues will be considered. First, the effect of the adaptation strategies for reducing task-dependent (TDEP) and speaker-dependent (SDEP) variability will be considered. Second, the effect of the applied phonotactic bigram constraints in decoding will be evaluated with respect to an unconstrained (zerogram) decoding. Third, the performance of the non-linear neural network (NN)-based mapping procedure will be presented.

The PV verification performance was found to vary across phoneme classes. For example, when the results in Table 6 are reported separately for phonemes classified as vowels and non-vowels, the performance for the vowel class is considerably worse than the non-vowel class. For the TDEP MAP adaptation case using the zerogram network, the vowel class EER is approximately 18 percent higher than the EER obtained for the non-vowel class. Vowels represent 44 percent of the total phoneme occurrences in the corpus. This surprising difference in EER is partly due to the human labeling strategy. Rather forgiving subjective judgments were made by the labelers when deciding whether a given utterance contained a "pronunciation variant" of a phoneme as opposed to a labeled mispronunciation error. This results in many cases where the decision threshold defines a phoneme instance to be mispronounced when the reference label indicates the phoneme was correctly pronounced. There is less ambigu-

ity in human labelers' decisions for the labeling of deletion errors. The higher EER observed for vowels results from the fact that substitution errors are more common for vowels and deletion errors are more common for non-vowels.

Table 8 presents the global PV performance for a variety of experimental conditions where each is delineated in the first column of the table. The second and third columns of Table 8 display the performance in EER for the zerogram and bigram recognition networks respectively. The results in Table 8 are obtained on a test set consisting of 2 394 utterances and 12 264 monophone test trials. These include 10 083 phonemes labeled as being correctly pronounced and 2 128 labeled as incorrectly pronounced. The 2 128 'incorrect' test trials correspond to phoneme instances that have been either mispronounced by the test speaker (substituted for another phoneme) or deleted altogether.

There are several observations that can be made from the results given in Table 8. First, from the first row of the table, it is clear that the EER for verifying phoneme-level pronunciation task-independent HMM models trained from the Albayzín speech corpus is fairly high. An EER of over 25 percent is obtained when no phonotactic constraints are applied in decoding. An EER of 22 percent is obtained when the bigram network is used. The second row of the table shows that MAP/MLLR adaptation of the HMM to the corpus of unimpaired children and young adults speaking utterances of the same vocabulary words results in approximately twenty percent decrease in EER. This rather significant improvement is due largely to the significant mismatch in speaker characteristics that exists between the largely adult speaker population in the Albayzín corpus and the unimpaired younger speaker population in the adaptation corpus.

The age of the children and young adults in the corpora used here ranged from 11 to 21 years old. The age of members of the speaker population in the Albayzín training corpus ranged from 19 to 64 years old. The ages of one third of the speakers in the Albayzín corpus were between 39 and 64, approximately two thirds of the speakers were between the ages of 23 and 38, and less than one percent of the speakers in the Albayzín corpus were less than 22. Hence, the degree of overlap between the ages of the two speaker populations was extremely small.

The third row of Table 8 shows that speaker-dependent MLLR adaptation of the TDEP HMM models using 57 utterances from each test speaker results in approximately seven percent decrease in EER. Note that the speaker-dependent adaptation data includes both correctly pronounced phonemes and phonemes that were mispronounced by the impaired speakers. Including the mispronounced phonemes in the adaptation data may limit the potential performance improvements that are achievable in this scenario.

| TDEP Verification Performance using Reduced Test Set (EER) | |
|---|---|
| TDEP MAP/MLLR Adaptation | 20.5% |
| TDEP + NN Mapping | 19.7% |

Table 9
Task-dependent (TDEP) phoneme detection performance using unconstrained zerogram network obtained with and without speaker-independent NN-based non-linear mapping

The fourth row of the table shows the effect on performance when the same utterances used for MLLR adaptation are instead used to train the NN-based mapping described in Section 6.3. This results in a substantial 18% reduction in EER with respect to the TDEP case. Finally, comparing the EER displayed in the second and third columns of Table 8, the bigram phonotactic constraints result in a reduction in EER rate between 7% and 12%.

The issue of the statistical significance of differences between measures based on false accept rates and false reject rates has been addressed in the literature (Bengio and Mariethoz, 2004; Delong et al., 1988). However, there is no significance test for these applications that has become widely accepted in the speech and language community, so the results of these significance tests can be difficult to interpret. By any test, one would assume that the EER difference of 1.3 percent shown for rows two and three of Table 8 are at best barely statistically significant. This equal error rate point corresponds to a difference of 130 false rejection trials out of 10083 correctly pronounced phonemes and 27 false acceptance trials out of 2128 incorrectly pronounced phonemes. Without computing confidence intervals on these outcomes, one cannot conclude with any certainty that the resulting estimate of the difference in error rates for this case is significant.

Figure 9 displays the pronunciation verification performance over the 2 394 utterance test set in the form of ROC curves. The ROC curves labeled TIND, TDEP, and SDEP in Figure 9 correspond to the systems whose zerogram EER results are given in rows two through four of Table 8. Note that the performance characteristics are well behaved in that the same rank order of performance is achieved by the three systems at all operating points.

While the NN-based nonlinear mapping was shown in Table 8 to provide a substantial reduction in EER, the scenario followed for the system in Table 8 involved using speaker-dependent data in training the NN. In order to investigate the effect of this mapping in a speaker-independent scenario, the 14-speaker training set was divided in half. Utterances from the first seven speakers were used for training the NN-based mapping and utterances from the second set of seven speakers were used as a test set. The results for this revised training and testing scenario on the reduced test set are displayed in

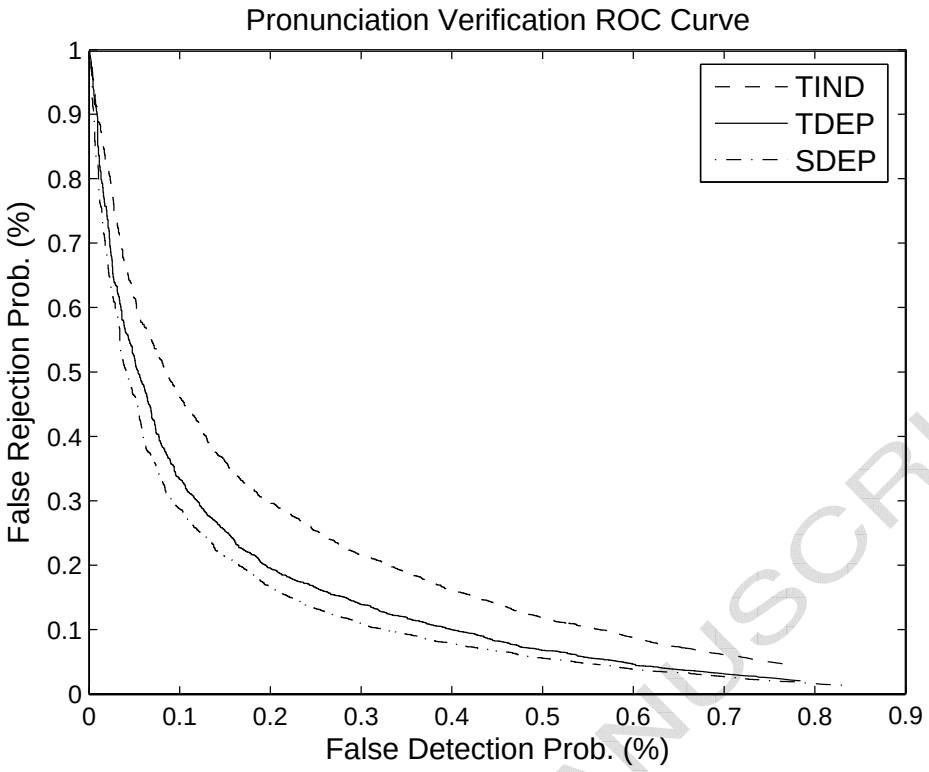


Fig. 9. ROC curves displaying phoneme-level verification using three different HMM models: task independent (TIND) models, task adapted (TDEP) models, and speaker adapted (SDEP) models. A zerogram, or unconstrained, ASR network was used to generate phoneme lattices

Table 9. It is clear from Table 9 that the impact of the NN-based mapping in reducing the EER relative to the TDEP performance is far less for the speaker-independent training of the NN than it is for the speaker-dependent case reported in Table 8. Although these last results have to be considered carefully due to the reduced test set used (only half of the sessions per speaker).

## 7 Discussion

In this section, it will be discussed how speech technologies can help in the improvement of pronunciation skills by young impaired children in the application "Vocaliza". The corpus used for the evaluation of these speech technologies (ASR and PV) gathers a group of children and young adults with speech therapy needs who are potential users of the "Vocaliza" tool. The baseline result in ASR (66.8% of WAC) indicates the strong loss of performance due to the speech disorders of the speakers. Hence, research in speech technologies in the presence of impaired speech is necessary to achieve a better performance in the proposed CASLT tools.

Regarding ASR, speaker adaptation provided a significant increase in WAC (63.83%), showing the strong influence of the speakers' disorders in the performance of the ASR system in "Vocaliza". Further studies assured this hypothesis, like the correlation between ASR results and the rate of mispronunciations per speaker. The ASR system is used in "Vocaliza" to accept or reject the speaker's utterance of a prompted word prior to evaluation. The increase in WAC achieved with speaker adaptation is necessary to avoid the frustration of the user when correctly pronouncing the word, while rejecting highly incorrect utterances. However, trying to evaluate whole words as correct or incorrect is influenced by the non-unique matching of the human-obtained phoneme labels to word labels, resulting in the need of phoneme-level PV for advanced evaluation.

In terms of PV, a major reduction in EER was achieved with a non-linear mapping to the phonetic labels made over the impaired speech corpus. This mapping makes the PV algorithm match the possible corrective scheme given by an educator in a speech therapy class. The final result in EER is promising for the use of this phoneme-level PV algorithm in tools requiring a phoneme-level measure of the articulation abilities of the user like second language learning tools (Rodríguez, 2008).

## 8 Conclusions

In this paper, a set of tools for speech and language therapy have been presented under the framework of "Comunica". These tools are oriented to children who need to train their phonatory, articulatory and descriptive abilities. Two of these tools make use of speech technologies like speech recognition and pronunciation verification to provide of real language improvement in the user. ASR can provide a simple but efficient word-level pronunciation verification while a novel PV system has been tested to detect mispronunciations at the phoneme-level. Speech technologies jointly with AAC systems and a simplified user interface allow the unsupervised automation of the process of speech and language therapy for children.

Future work might include studies in ASR and PV considering an open-vocabulary task. This study would validate tools for teaching language understanding like "Cuéntame" in the same way that it has been done with "Vocaliza" in this paper. An open-vocabulary domain would have to consider the study in language modeling for language impaired individuals. The treatment of OOV words in the ASR system would also be required in this domain.

## 9 Acknowledgments

The authors want to acknowledge José Manuel Marcos, César Canalís, Pedro Pegero and Beatriz Martínez from the School for Special Education "Alborada", located in Zaragoza (Spain), for their collaboration in this work.

Special thanks to Antonio Escartín for his work and to Victoria Rodríguez from the Vienna International School (Austria) for her interest in this work.

## References

Acero-Villán, P., Gomis-Cañete, M.-J., 2005. Tratamiento de la Voz (Manual Práctico). Ed. CEPE, Madrid, Spain.

Aguinaga, G., Armendia, M., Fraile, A., Olangua, P., Uriz, N., 2004. Prueba del Lenguaje Oral Navarra - revisada. TEA S.A., Madrid, Spain.

Alarcos, E., 1950. Fonología Española. Ed. Gredos, Madrid, Spain.

Albor, J.-C., 1991. ELA - Examen Logopédico de Articulación. Ed. CEPE, Madrid, Spain.

Bengio, S., Mariethoz, J., June 2004. A statistical significance test for person authentication. In: Proceedings of the Speaker and Language Recognition Workshop ODYSSEY 2004. Toledo, Spain, pp. 237–244.

Coorman, G., Fackrell, J., Rutten, P., Van Coile, B., October 2000. Segment selection in the L&H RealSpeak laboratory TTS System. In: Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP-Interspeech). Beijing, China, pp. 395–398.

Cucchiarini, C., Neri, A., de Wet, F., Strik, H., September 2007. ASR-based pronunciation training: scoring accuracy and pedagogical effectiveness of a system for Dutch L2 learning. In: Proceedings of the 10th European Conference on Speech Communication and Technology (Eurospeech-Interspeech). Antwerp, Belgium, pp. 2181–2184.

Deller, J.-R., Hsu, D., 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.

Delong, E.-R., Delong, D.-M., Clarke-Pearson, D.-L., September 1988. Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Journal of Biometrics 3, 837–845.

Dempster, A.-P., Laird, N.-M., Rubin, D.-B., 1977. Maximum Likelihood for incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39 (1), 1–21.

Duchateau, J., Cleuren, L., Van Hamme, H., Ghesquiere, P., September 2007. Automatic assessment of childrens reading level. In: Proceedings of

the 10th European Conference on Speech Communication and Technology (Eurospeech-Interspeech). Antwerp, Belgium, pp. 1210–1213.

Escartín, A., Saz, O., Vaquero, C., Rodríguez, W.-R., Lleida, E., 2008. "Comunica" framework web site.
URL `http://www.vocaliza.es`

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., Hohmann, S., September 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). Budapest, Hungary, pp. 1067–1070.

Gauvain, J.-L., 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, M., Narayanan, S., April 2008. Investigating assessment of reading comprehension in young children. In: Proceedings of the 2008 International Conference on Acoustics, Speech and Signal Processing (ICASSP). Las Vegas (NV), USA, pp. 5057–5060.

Goronzy, S., Kompe, R., 1999. A combined MAP + MLLR approach for speaker adaptation. In: Proceedings of the the Sony Research Forum 99 1, 9–14.

Granstroem, B., September 2005. Speech technology for language training and e-inclusion. In: Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech-Interspeech). Lisbon, Portugal, pp. 449–452.

Hatzis, A., October 1999. Optical Logo-Therapy: Computer-based audio-visual feedback using interactive visual displays for speech training. Ph.D. thesis, Department of Computer Science, The University of Sheffield, Sheffield, United Kingdom.

Hawley, M., Enderby, P., Green, P., Brownsell, S., Hatzis, A., Parker, M., Carmichael, J., Cunningham, S., O'Neill, P., Palmer, R., August 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.

Ito, A., Tsutsui, R., Makino, S., Suzuki, M., September 2008. Recognition of English utterances with grammatical and lexical mistakes for dialogue-based CALL system. In: Proceedings of the 11th International Conference on Speech Communication and Technology (ICSLP-Interspeech). Brisbane, Australia, pp. 2819–2822.

Justo, R., Saz, O., Guijarrubia, V., Miguel, A., Torres, M.-I., Lleida, E., February 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 (QC), Canada.

Kim H., Hasegawa-Johnson M., Perlman A., Gunderson J., Huang T., Watkin K., Frame S., September 2008. Dysarthric Speech Database for Universal Access Research. In: Proceedings of the 2008 International Conference on Spoken Language Processing (ICSLP - Interspeech). Brisbane (QLD), Australia, pp. 1741–1744.

Koehn, P., September 2005. Europarl: A parallel corpus for statistical machine translation. In: Proceedings of the 10th Machine Translation Summit. Phuket,Thailand.

Kornilov, A.-U., September 2004. The Biofeedback program for speech rehabilitation of oncological patients after full larynx removal surgical treatment. In: Proceedings of the 9th International Conference on Speech and Computer (SPECOM). St. Petersburg, Russia.

Lefévre, J.-P., 1996. Harp: An autonomous rehabilitation system for hearing impaired people. Tech. rep., TIDE Project 1060.

Legetter, C.-J., Woodland, P.-C., 1995. Maximum Likelihood Linear Regression for speaker adaptation of the parameters of continuous density Hidden Markov Models. Computer Speech and Language 9, 171–185.

Lleida, E., Rose, R.-C., March 2000. Utterance Verification in continuous speech recognition: Decoding and training procedures. IEEE Transactions on Speech and Audio Processing 8 (2), 126–139.

Mangu L., Brill, E., Atocke, A., September 1999. Finding Consensus Among Words: Lattice-Based Word Error Minimization. In: Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech-Interspeech). Budapest, Hungary pp. 495–498.

Martínez, B., Peguero, P., Ezpeleta, J., Falcó, J., Lleida, E., Mínguez, J., Saz, O., November 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.

Menéndez-Pidal, X., Polikoff, J.-B., Peters, S.-M., Lorenzo, J., Bunnell, H.-T., October 1996. The Nemours database of dysarthric speech. In: Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP-Interspeech). Philadelphia (PA), USA, pp. 1962–1965.

Monfort, M., Juárez-Sánchez, A., 1989. Registro Fonológico Inducido (Tarjetas Gráficas). Ed. Cepe, Madrid, Spain.

Monfort, M., Monfort-Juárez, I., 2001. En La Mente: Un soporte gráfico para el entrenamiento de las habilidades pragmáticas en el niño. Entha ediciones, Madrid, Spain.

Moreno, A., Lindberg, B., Draxler, C., Richard, G., Choukri, K., Euler, S., Allen, J., June 2000. Speech Dat Car: A large speech database for automotive environments. In: Proceedings of the II Language Resources European Conference. Athens, Greece, pp. 895–900.

Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J.-B., Nadeu, C., September 1993. Albayzín speech database: Design of the phonetic corpus. In: Proceedings of the 3th European Conference on Speech

Communication and Technology (Eurospeech-Interspeech). Berlin, Germany, pp. 653–656.

Navarro-Mesa, J.-L., Quintana-Morales, P., Pérez-Castellano, I., Yáñez, J. E., May 2005. Oral corpus of the project HACRO (Help tool for the confidence of oral utterances). Tech. rep., Department of Signal and Communications, University of Las Palmas de Gran Canaria.

Oester, A.-M., House, D., Protopapas, A., Hatzis, A., May 2002. 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, pp. 45–48.

Patel, R., March 2002. Phonatory control in adults with cerebral palsy and severe dysarthria. AAC Augmentative and Alternative Communication 18, 2–11.

Pratt, S.-R., Heintzelman, A.-T., Deming, S.-E., October 1993. The efficacy of using the IBM Speech Viewer vowel accuracy module to treat young children with hearing impairment. Journal of Speech and Hearing Research 36 1063–1074.

Rabiner, L.-R., Schafer, R.-W., 1978. Digital processing of speech signals. Prentice-Hall (Signal Processing Series), Englewood Cliffs (NJ), USA.

Rodríguez, W.-R., Vaquero, C., Saz, O., Lleida, E., June 2008. Speech technology applied to children with speech disorders. In: Proceedings of the 4th Kuala Lumpur International Conference on Biomedical Engineering. Kuala Lumpur, Malaysia.

Rodríguez, V., November 2008. El uso de herramientas multimedia para la práctica de la pronunciación en clases de ELE con adolescentes. Master Thesis in Enseñanza del Español como Lengua Extranjera. Department of Applied Languages, University Antonio de Nebrija, Madrid, Spain.

Sanders, E., Ruiter, M., Beijer, L., Strik, H., September 2002. Automatic recognition of dutch dysarthric speech: A pilot study. In: Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP-Interspeech). Denver (CO), USA, pp. 661–664.

Saz, O., Miguel, A., Lleida, E., Ortega, A., Buera, L., September 2006. 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). Pittsburgh (PA), USA, pp. 993–996.

Saz, O., Rodríguez W.-R., Lleida, E., Vaquero C., October 2008. A novel corpus of children disordered speech. In: Proceedings of the First Workshop on Child, Computer and Interaction. Chania, Greece.

Vaquero, C., Saz, O., Lleida, E., Rodríguez, W.-R., April 2008. E-inclusion technologies for the speech handicapped. In: Proceedings of the 2008 International Conference on Acoustics, Speech and Signal Processing (ICASSP). Las Vegas (NV), USA, pp. 4509–4512

Vicsi, K., Roach, P., Oester, A., Kacic, Z., Barczikay, P., Sinka, I., September 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). Budapest, Hungary, pp. 859–862.

Yin S.-C., Rose, R.-C., Saz, O., Lleida, E., April 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). Taipei, Taiwan.

Zhang, F., Huang, C., Soong, F. K., Chu, M., Wang, R., April 2008. Automatic mispronunciation detection for Mandarin. In: Proceedings of the 2008 International Conference on Acoustics, Speech and Signal Processing (ICASSP). Las Vegas (NV), USA, pp. 5077–5080.