E-INCLUSION TECHNOLOGIES FOR THE SPEECH HANDICAPPED

Carlos Vaquero, Oscar Saz, Eduardo Lleida and W.-Ricardo Rodríguez

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

ABSTRACT

This paper addresses the problem that disabled people face when accessing the new systems and technologies that are available nowadays. The use of speech technologies, specially helpful for motor handicapped people, becomes unapproachable when these people also suffer speech impairments, making the gap in the society wider for them. As a way to include speech impaired people in the technological society of today, two lines of work have been carried out.

On one hand, a computer-aided speech therapy software has been developed for the speech training of children with different disabilities. This tool, available for free distribution, makes use of different state-of-the-art speech technologies to train different levels of the language. As a result of this work, the software is being used currently in several centers for special education with a very encouraging feedback about the capabilities of the system.

On the other hand, research on the use of Automatic Speech Recognition (ASR) systems for the speech impaired has been carried out. This work has focused on current techniques of speaker adaptation to know how these techniques, fruitfully used in other tasks, can deal with this specific kind of speech. The use of Maximum A Posterior (MAP) obtains an improvement of 60.61% compared to the results of a baseline speaker independent model.

Index Terms— Handicapped aids, User interfaces, Speech processing, Speech recognition

1. INTRODUCTION

During last years, technological development has changed our way of life. Today we are used to have access to many tools in our daily routine which were rare or even non-existent some years ago. This tools (computers, laptops, mobile phones and so on) make us more efficient in our work and help us to make the most of our free time. However, all this technological development is enlarging the barrier that handicapped people must overcome to have the same possibilities than non-disabled people. Hence, the concept of e-inclusion arises, as the need of including handicapped people into the technological world of today.

The use of speech technologies is a clear example of this growing technological gap for handicapped people. Dictation systems or environment control systems, that for us are starting to become more common and help us to make our life simpler, could be a great help for people with motor disabilities, that could compensate their physical impairments with the use of their speech. But many of these people also present speech impairments, and, nowadays, speech technology systems are not still ready to face successfully these kind of situations outside controlled environment with a collaborative unimpaired user.

Two ways of solving this gap are then available. On one hand, any kind of improvement in the techniques for speech therapy will make the patients improve their quality of speech and being more likely to use all of these speech technology systems. Specifically, the development of computer-aided speech therapy tools is a way of helping speech therapists in their work. On the other hand, researching techniques to make Automatic Speech Recognition (ASR) systems more robust against impaired speech will help for the inclusion of the speech handicapped in these technological advances.

Regarding the development of computer-aided speech therapy tools, many European projects related to speech technology and speech therapy such as Orto Logo-Paedia [1], SPECO [2], ISAEUS [3] and HARP [4] have been carried out during the last decade, some of them resulting in the development of software applications for speech therapy. However, there are no versions of these softwares available in Spanish language, so the applications developed in these projects can not be used by patients and speech therapists to train communication skills in this language.

Due to that, the Aragon Institute for Engineering Research (I3A) with the collaboration of experts in pedagogy and speech therapy from the Public School for Special Education (CPEE) "Alborada" has developed a research work which aims for providing speech technologies as a tool to aid speech impaired and handicapped people. This article, which explains the work carried out, is organized as follows: Section 2 describes "Vocaliza": a freeware application for speech therapy in Spanish language, developed within this work. Section 3 analyses speaker adaptation as a way to make the use of speech technologies easier for speech impaired people, and finally, section 4 shows the conclusions of this work.

2. VOCALIZA

Vocaliza is a freeware application based on speech technologies aimed to help speech impaired people to improve their communication skills and to provide handicapped people the use of speech technologies as a way of improving their quality of life.

For this purpose, the collaboration of experts in speech therapy and pedagogy is strongly necessary. The assistance of the staff of the CPEE "Alborada", located in Zaragoza, Spain, which is the Reference Centre for Technical Aids and Communication as appointed by the Regional Government of Aragón was essential for setting the application requirements prior to the start of the work and for reaching the objectives of this work, as they had been tracking the whole development process.

This work has been supported by the national project TIN-2005-08660-C04-01 from MEC of the Spanish government.

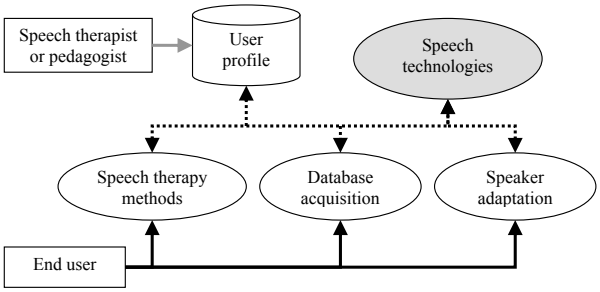

Fig. 1. Vocaliza block diagram.

The final architecture of the application is represented as a block diagram in Figure 1. Blocks exchange audio information (solid black arrow), user information (black dotted arrow) and configuration information (solid grey arrow). As shown in Figure 1, the application must be configured previously by a pedagogist or a speech therapist, to obtain the desired operation. After that, the end user, a speech impaired person, will be able to use the application with little or no supervision.

Every block functionality is explained next.

2.1. Speech Therapy

One of the purposes of the application is providing methods for improving the communication skills of the user. The application exercises three levels of language, namely the phonological, semantic and syntactic levels. Each level is exercised by a different method which is shown as a game, in order to attract young users.

Phonological level is exercised encouraging the user to utter a set of words previously selected by a speech therapist or pedagogist during the configuration procedure. These words are selected to focus on the user speech pathology. The application evaluates every utterance and displays a grade with an animated motion on the screen, that the user will be able to understand easily.

Semantic level is exercised by means of a set of riddles, previously defined by a speech therapist or pedagogist. The application asks a question to the user and gives three possible answers. The user must utter the correct answer 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.

Syntactic level is exercised encouraging the user to utter a set of sentences, previously selected by a speech therapist or pedagogist. Again, the application will evaluate user utterances to display a grade, marking the improvement of the user.

All games are based on ASR, which will decide if the word or sentence uttered by the user is the one the application was expecting.

2.2. Database acquisition

The application provides a method to record user utterances, designed to attract young users, which will make easier to obtain databases of impaired speech. This method is quite simple: first of all, a set of words and sentences to record is defined. Each word and sentence will be shown on the screen as an image or set of images which will come with associated sounds (usually a correct utterance of the word/sentence), and the user will have to utter it. In order to obtain high quality databases, supervision during the procedure is strongly recommended.

2.3. Speaker adaptation

Speaker adaptation techniques enable ASR systems to obtain better performance when used by a specific speaker. Vocaliza enables the user to make use of speaker adaptation in order to obtain acoustic models adapted to his/her speech impairments. The application will use these acoustic models every time the user accesses the speech therapy functionality. This will allow the speech impaired people who suffer serious disorders to use the application, since ASR systems may fail when used by people suffering severe speech disorders. Moreover, since acoustic models contain information about the impairment of the user, they will allow the application to track improvements in the communication skills of the user, and to finally show a grade in every game capturing the user improvements.

In addition, speaker dependent acoustic models can be very useful for handicapped people. In fact, these models will improve ASR performance in environmental management systems which will make their life easier.

2.4. Speech technologies

Vocaliza is based on four different speech technologies, which are ASR, Text-To-Speech (TTS), speaker adaptation and Utterance Verification (UV). Their functionality within the application is explained next.

ASR constitutes the core of the application. Speech therapy games need 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.

TTS provides a way of showing the user the correct pronunciation of a word or sentence. This is a useful feature in speech therapy games as well as in database acquisition procedure. As soon as a speech therapist or pedagogist adds a new word, sentence or riddle to the application, the TTS system generates the correct Spanish utterance of the corresponding word, sentence or question. However, TTS may be a very strict method to teach the user how to pronounce a word or a sentence. Thus, to provide flexibility, Vocaliza allows speech therapists to record word, riddle, and sentence utterances, that the application will use instead of TTS. This way, different utterances are shown depending on the age, the speech impairments or other requirements of the user.

Speaker adaptation enables the application to estimate a speaker dependent acoustic model for every user. Vocaliza uses Maximum A Posteriori (MAP) estimation [5] which, given a speaker independent acoustic model and a set of user utterances, can estimate a speaker dependent acoustic model. MAP is a well known and reliable estimation method which does not require a great number of utterances to retrieve a reliable speaker dependent acoustic model. This is a very interesting feature since the application will estimate acoustic models from a set of utterances recorded by the own user, which in most cases will consist of a small number of utterances due to two factors: speech therapists can not spend long time recording speech of every user, and users suffering speech impairments will find very hard and tiring to record a great amount of utterances. Moreover, MAP estimation convergence make this method a very interesting one when the number of utterances is a priori unknown.

Utterance Verification (UV) is a technique embedded in the application to provide a mechanism to evaluate the improvement of user communication skills. Vocaliza uses a Likelihood Ratio (LR) based UV [6] procedure to assign a measure of confidence to each hypothesized word in an utterance. This procedure gives the confidence measure as the ratio of the target hypothesis acoustic model

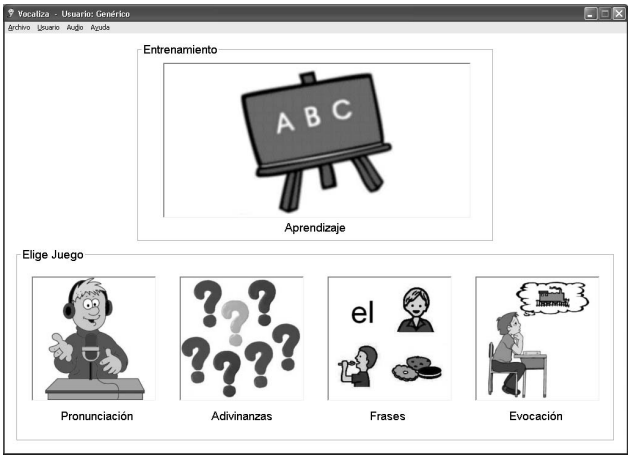

Fig. 2. Vocaliza main window.

likelihood versus an alternate hypothesis acoustic model likelihood. Choosing suitable acoustic models as target and alternate hypothesis can provide a measure of confidence which quantifies improvement in user speech. To achieve this, the application uses a speaker independent acoustic model, which is assumed to model correct speech, as target hypothesis, and a speaker dependent acoustic model, which is assumed to be adapted to the speech disorders of the user, as alternate hypothesis. Therefore, this measure of confidence involves a relative evaluation method to quantify the improvement of the communication skills of the user.

2.5. User profile

User profile stores all information regarding user configuration, including all words, riddles and sentences selected by a speech therapist to exercise the speech of the user. It also stores all utterances recorded by the user and all speaker dependent acoustic models. This provides flexibility so that speech therapists will be able to work with different patients in a fast and easy way by merely loading the user profile in the application.

2.6. User interface

Usability is a very important feature to take into account in this kind of application, since all functionality provided would not be successful if speech therapists, pedagogist and, above all, speech impaired and handicapped people find it difficult to use the application. In order to assure usability, it is important to design an easy to use interface which in addition is attractive to young users. Vocaliza achieve this by using cartoons to represent every game, and confining all configuration options and advanced functionalities in the main menu. Fig. 2 shows Vocaliza main window distribution.

Moreover, end users do not have to know how to use a computer to exercise their communication skills with the application. Speech therapy games work only with audio input, so users will only speak to a microphone. Once launched, speech therapy games will wait until the user utters the requested word or sentence to go on, and will not request any kind of input different from speech (keyboard or mouse) but to finish the game.

In addition, this way of working encourages speech impaired people to use speech in their lives, which will help to improve their communication skills, as they realize that they can achieve goals with their speech.

3. ASR FOR IMPAIRED USERS

Being ASR the core of the "Vocaliza" software and having been developed with the aim of providing speaker adaptation and ASR to the speech impaired, the research in this field is clearly needed. Even when some studies have been made [7], the difficulties of this task has made unable to get a real good performance.

In this work, one of the aims is measuring the possibilities of using ASR systems to improve the quality of life of disabled people. With that purpose, a database with impaired children speech was used to obtain a baseline of the performance of an ASR system with this kind of users.

3.1. Impaired speech database

An added difficulty arises when working with impaired speech in the Spanish language, this is the fact that there are no relevant databases of impaired speech in this language. Research in speech technologies often make use of large databases to train and develop acoustic and language models in order to model human speech, but there is a lack of databases when working in some specific tasks like this one.

For this reason, this work also comprehends the recording of a database containing speech uttered by children with different impairments. This database contains the speech of 14 children (7 girls and 7 boys) attending the CPEE "Alborada" . Their ages range from 11 to 21 years old and their disabilities comprehend physical and psychical handicaps like cerebral palsy, Down's syndrome and similar impairments that affect their language in all the levels: from the phonological level to the morphological, syntantic and/or semantic levels.

The set of words chosen for the recordings is the Induced Phonological Register (RFI) [8]. These 57 words, arranged in terms of their difficulty is a very well-known set of words for the speech therapists in Spain. The set of words contains a rich selection of all the Spanish phonemes in several situations of phonemes boundaries and neighborhood relations.

The utterances were recorded in the facilities of the CPEE "Alborada" , under the supervision of at least one of the members of the Communication Technologies Group (GTC) and the surveillance of at least one of the staff members of the CPEE "Alborada". Every speaker recorded 4 series of the RFI in 4 different days to avoid the tiredness of the speaker during the process and also to capture the intra-speaker variability as another factor in the database. This way, a total number of 3,192 isolated word utterances are into the database.

The database was recorded with a close-talk wireless microphone (AKG C444L). The use of a close-talk microphone gave a noise-free recordings, where the average Signal to Noise Ratio is 26.35 dB. The use of a wireless microphone was to avoid that the children were attached to the recording system, so they could feel more comfortable during the recording sessions. Signals were recorded with a 16 kHz sampling frequency and a depth of 16 bits.

3.2. Baseline and speaker independent models

The ASR system in this work uses a 39 Mel Frequency Cepstral Coefficient (MFCC) parametrization, with 12 static parameters, 12 delta parameters, 12 delta-delta parameters and the log-energy plus its first and second derivative. The signals were windowed with a Hamming window of 25 ms. length, with an overlap of 15 ms. An HMM-based speaker independent acoustic model was used with a set of 744 context-dependent units. This model also included two units to model begin-end silence and interword silence. All units

Speaker	#01	#02	#03	#04	#05
WER	15.79%	35.96%	22.37%	12.72%	*62.72%*
Speaker	#06	#07	#08	#09	#10
WER	4.82%	33.33%	40.35%	32.46%	41.67%
Speaker	#11	#12	#13	#14	
WER	12.28%	*67.98%*	*83.33%*	27.63%	

Table 1. *Baseline WER results.*

Speaker	#01	#02	#03	#04	#05
WER	0.88%	10.53%	0.44%	0.44%	*45.62%*
Speaker	#06	#07	#08	#09	#10
WER	0.00%	7.46%	24.12%	10.09%	10.97%
Speaker	#11	#12	#13	#14	
WER	0.88%	*14.04%*	*64.47%*	4.39%	

Table 2. *Speaker dependent WER results.*

were modeled as 1-state units with 32 gaussians per state. This speaker independent model was trained with the 44108 noise-free signals of the adult Spanish speech databases SpeechDat-Car, Albayzin and Domolab.

The baseline results in terms of Word Error Rate (WER) are shown in Table 1. The average WER for the 14 speakers (#01 to #14) is 35.24%, ranging from 4.82% for the speaker #06 to 83.33% for the speaker #13. This speaker, like #05 and #12 have a WER over 50%, which clearly shows the severeness of their speech impairment. It is also noticeable the wide difference among different speakers, as all of them suffer from different kinds of impairments.

3.3. Results with speaker dependent models

The most simple and useful way in many cases to improve the performance of an ASR system is the use of speaker adaptation to create speaker dependent models that reduce the mismatch due to the speaker variability. Algorithms like Maximum A Posterior (MAP) [5] or Maximum Likelihood Linear Regression (MLLR) [9] are widely used for this task and obtain good performance in many cases.

To check the possibility of using speaker adaptation as a way to make ASR systems available for speech impaired people some experiments with MAP adaptation were made. A leave-one-out strategy was used where three of the recorded series for every speaker were used to create a speaker dependent model to obtain the ASR results over the remaining serie. Final results per speaker were obtained as the averaged sum of the four results obtained over the four series of the RFI that every speaker recorded for the database used in the work. This means that four speaker dependent models were trained for every one of the 14 speakers.

The results, as shown in Table 2, show an average WER of 13.88% (60.61% of improvement over the speaker independent results). It is noticeable that 6 out of the 14 speakers obtain a WER under 5%, while the speakers with the most severe impairments (#13, #12 and #5) still keep a very high WER (64.47%, 45.62 and 14.04% respectively). This shows the need of research in other lines of improvement the acoustic modeling, like the addition of a lexicon model that characterizes the speech disorders of the speaker.

4. CONCLUSIONS

As a result of this work, a totally functional application which aims to help speech impaired people, providing speech therapy methods working in three levels of the language (phonological, semantic and syntactic), a friendly database acquisition interface and speaker adaptation has been developed. Although there is no objective data about usefulness and usability of the application available, it has been installed in several schools for special education, where teachers and speech therapists are using it for their daily work. They all have found the application easy to use, friendly and useful for their work, providing very interesting and encouraging feedback to keep us working in this research line.

In addition, a database containing utterances from speech impaired people has been acquired. Since there are not databases of this kind in Spanish language, this database will be very useful for present and future researching in ASR for speech impaired people.

In fact, this database has made possible to study speaker adaptation performance when used by speech impaired people. The obtained results show that speaker adaptation can be very useful to enable speech and motor handicapped people to use ASR in order to improve their quality of life, if they do not suffer a very serious speech impairment. However, in case of people suffering severe speech impairments, although speaker adaptation will improve ASR performance, it will not be enough to assure desired ASR operation. So there is still work to do in this research line.

5. REFERENCES

[1] Oester A-M, House D., Protopapas A., and Hatzis A., "Presentation of a new eu project for speech therapy: Olp (ortho-logopaedia)," in *Fonetik*, 2002.

[2] Vicsi K., Roach P., Oester A., Kacic Z., Barczikay Z., and Sinka I., "Speco — a multimedia multilingual teaching and training system for speech handicapped children," in *Eurospeech, 6th Conference on Speech Communication and Technology, Interspeech*, 1999.

[3] García Gómez et al., "Isaeus — speech training for deaf and hearing-impaired people," in *Eurospeech, 6th Conference on Speech Communication and Technology, Interspeech*, 1999.

[4] "Harp – an autonomous speech rehabilitation system for hearing impaired people," Final report, HARP (TIDE project 1060), May 1996.

[5] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[6] E. Lleida and R.C. Rose, "Utterance verification in continuous speech recognition: Decoding and training procedures," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 126–139, 2000.

[7] J. R. Deller, D. Hsu, and L.J. Ferrier, "On the use of hidden markov modelling for recognition of dysarthric speech," *Computer Methods and Programs in Biomedicine*, vol. 35, pp. 125–139, 1991.

[8] M. Monfort and A. Juárez Sánchez, *Registro Fonológico Inducido (Tarjetas Gráficas)*, Cepe, Madrid, 1989.

[9] C.-J. Legetter and P.-C. Woodland, "Maximum likelihood linear regression for speaker adaptation of the parameters of continous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.