HUMAN LANGUAGE TECHNOLOGIES FOR SPEECH THERAPY IN SPANISH LANGUAGE

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

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

Abstract

This paper introduces Vocaliza, an application for computer-aided speech therapy in Spanish language based on the use of Human Language Technologies (HLT). The objective of this application is to help the daily work of the speech therapists that train the linguistic skills of Spanish speakers with different speech impairments, working at three levels of language: phonological, semantic and syntactic. Furthermore, Vocaliza is designed to enable those who suffer speech disorders to train their communication capabilities in an easy and entertaining way, with little or no supervision once a speech therapist has configured the application for the impairment of the user.

The HLT systems used in the application are Automatic Speech Recognition (ASR), speech synthesis, speaker adaptation and utterance verification. The ability of these technologies, namely ASR and speaker adaptation, to actually help users to improve their language is shown by means of the accuracy of the ASR system to detect correct and incorrect utterances according to a manual labeling of a recently acquired database containing impaired speech. The results show that accuracy reaches 87.66% when using speaker adaptation, due to its ability to model the inter speaker variability of every speaker but not their pronunciation errors.

Index Terms: Human Language Technology, speech therapy, Spanish language

1. Introduction

Recently, the demand for computer-aided speech therapy software has increased as computer technologies were getting more reliable and affordable to speech therapists and people suffering speech impairments. The most popular of these systems has been SpeechViewer by IBM, but the non existence of a version for the Spanish language and its lack of modularity made it very uncomfortable for speech therapists in Spain to use on a regular basis.

In terms of research work, during the last decade many European projects related to Human Language Technology (HLT) and speech therapy such as Orto Logo-Paedia [1], SPECO [2], ISAEUS [3] and HARP [4] have been carried out, some of them resulting in the development of software applications for speech therapy at the end of the research process. However, there are no versions of these softwares available in Spanish language, so the applications developed in these projects can not be used by speakers 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 Alborada has developed a research work which aims to provide speech technologies as a tool to aid speech impaired and handicapped people, obtaining as a result a software application for speech therapy in Spanish language, which is free to distribute. This article, which explains the work carried out to obtain the application, is organized as follows: section 2 describes the objectives that are set to the development of a computer-aided speech therapy software in Spanish language. In section 3, there is a wide description of the application architecture while section 4 explains the experiments and results carried out to validate the application. Finally the conclusions to this work are explained in section 5.

2. Objectives and Requirements

The objective of this work was the development of a free distribution software application for speech therapy in Spanish language.

For this purpose, the collaboration of experts in speech therapy and pedagogy is strongly necessary. This work has counted on the assistance of the staff of the Public School for Special Education Alborada, located in Zaragoza, Spain, which is a Reference Center for Technical Aids and Communication appointed by the Regional Government of Aragon. Their knowledge in different fields of work with disabled children 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 application development process.

The requirements set for the application can be separated from four points of view:

- In terms of linguistic levels, the application should train several levels of language, from phonological level to semantic and syntactic levels, in order to work on a wide range of speech imparements.
- Regarding application usability, the application should provide enough flexibility for speech therapists to work on different speech impairments, while methods used to treat these impairments should be amusing to attract end users (mainly children).
- The application should have a modular way of dealing with the users, this is, information about every user speech impairments and most suitable methods to train user speech should be stored in order to enable speech therapist to work with different users in an easy way.
- The application should be easy to use, as speech therapists and speech impaired people may not be used to work with computers.

All these requirements were taken into account for the final development of the application, whose given name was Vocaliza.

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

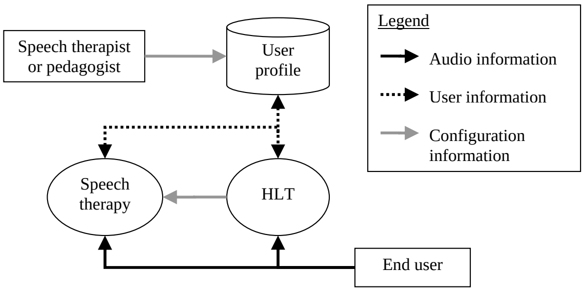

Figure 1: *Block diagram of Vocaliza.*

3. Vocaliza Architecture

Vocaliza architecture can be summed up in a block diagram as shown in Fig. 1. Blocks exchange audio information (solid black arrow), user information (dotted black arrow) and configuration information (solid grey arrow). As shown in Figure 1, the application must be configured previously by a speech therapist, to obtain the desired operation, and after that, the end user, which will be a speech impaired person, will be able to use the application with little or without supervision. Every block functionality is explained next.

3.1. Speech Therapy

The main purpose of Vocaliza is to provide methods for improving user communication skills. The application trains three levels of language, namely phonological, syntactic and semantic levels. Each level is trained by a different method which is shown as a game, in order to attract young users.

Phonological level is trained forcing the user to utter a set of words previously selected by a speech therapist during a configuration procedure. These words are selected to focus on every user specific speech impairment. The application evaluates every utterance and displays a mark with an animated motion on the screen, that the user will be able to understand easily.

Syntactic level is trained forcing the user to utter a set of sentences, previously selected by a speech therapist. Again, the application will evaluate user utterances to display a mark, showing user improvement.

Semantic level is trained by means of a set of riddles, previously defined by a speech therapist. The application ask 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 mark depending on the user ability to solve the riddle.

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

Fig. 2 shows a screen shot of the main window of Vocaliza. In this window, every game is represented as a picture in order to enable the user to access the desired game easily.

3.2. User Profile

User profile stores all information regarding user configuration, including all words, riddles and sentences selected by a speech therapist to train user speech, as well as all utterances recorded by the user or all speaker dependent acoustic models. This provides flexibility and modularity so that speech therapists will be able to work with different patients fast and easily, merely loading the user profile in the application.

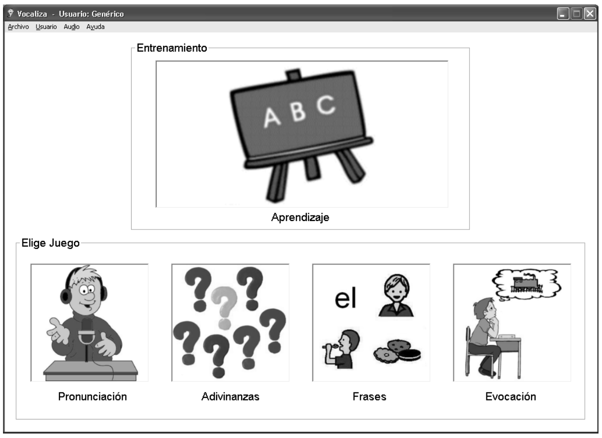

Figure 2: *Main window of Vocaliza.*

3.3. HLT in Vocaliza

Most of Vocaliza functionalities are provided by different HLTs, which are 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.

The ASR system used is based on Hidden Markov Models (HMM). Speech signals are acquired with a sampling frequency of 16 kHz and a bit depth of 16 bits. Signals are windowed with a Hamming window of 25 ms length, with an overlap of 15 ms, and the features used for the ASR are 37 MFCC (Mel Frequency Cepstral Coefícients), consisting on 12 static parameters, 12 delta parameters, 12 delta-delta parameters and the delta-logenergy. The acoustic model is composed by a set of 822 context dependent units plus a silence model and an interword model for a total set of 824 units. Every unit is modeled with 1 state per model and a 16-Gaussian mixture for every state.

Speech synthesis provides a way to show the user how a word or sentence should be pronounced, which is useful in speech therapy games. As soon as a speech therapist adds a new word, sentence or riddle to the application, it is able to synthesize a correct Spanish utterance of the corresponding word, sentence or question. However, speech synthesis 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, which the application will use instead of speech synthesis, in order to show different utterances depending on user age, speech impairments and other requirements of the user.

Speaker adaptation enables the application to estimate speaker dependent acoustic models adapted to each 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, adapted to the user. MAP is a well known and reliable estimation method which does not require a great number of utterances to retrieve a reliable acoustic model adapted to the user. This is a very interesting feature since the application will estimate acoustic models from a set of utterances recorded by the 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 with speech impairments will find very hard and tiring to record a great amount of speech utterances. Moreover, MAP estimation convergence make this method a very interesting one when the number of utterances is a priori unknown.

Speaker adaptation is strongly necessary in this application since impaired speech can reduce dramatically ASR systems performance, so that users suffering severe speech impairments would not be able to train their speech with this application using speaker independent models.

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 likelihood with respect to 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 impaired speech, as alternate hypothesis. Therefore, this measure of confidence involves a relative evaluation method to quantify improvement of user communication skills.

4. Evaluation and Results

To evaluate the performance of the HLT used in this application and how it can be used to improve the language abilities of the user, a set of recordings in an actual environment of use were made for testing.

4.1. Database acquisition

These recordings form a database of impaired speech recorded from 14 young speakers ranging from 11 to 21 years old, 7 boys and 7 girls. The recordings were made in the same school facilities they are attending currently, and acquired via a wireless close-talk microphone connected to a laptop where the audio capture feature of Vocaliza was used to store the signals. The Signal-to-Noise Ratio (SNR) in the signals is 26.35 dBs, an optimal value for the correct operation of the application and for the evaluation of the system. The 14 speakers suffer from different physical and mental disabilities that affect in several ways their speech and language abilities.

The vocabulary used for the recordings was a set of 57 words gathered in the "Registro Fonológico Inducido" (RFI) [7]. This set of words is a common tool in the community of speech therapy in Spain for the diagnosis of speech disorders as it contains all the phonemes and most of the allophones in Spanish language as well as different combinations of them. The average length of the words is 5,22 phonemes per word. Every speaker recorded four sessions and uttered these 57 words once in every session. Hence, the total number of utterances of isolated words in the database is finally 3,192 utterances. Sessions were recorded on different days to be more realistic with the presence of intra speaker variability among sessions.

During the recordings of the impaired speech database, another database containing speech from children in the range from 11 to 18 years old without any kind of disability was recorded. This database stores speech from 168 speakers to be used as a reference of the speech in the age range of the impaired speakers. Every speaker recorded one session of the 57 words leading to a total number of utterances of 9,576. The same recording process was used for this database and the average SNR is 25.59 dBs.

Training	ML	MAPref	MAPspk
WER	52.22%	31.45%	16.07%

Table 1: *WER results for different acoustic models. ML stands for the adult acoustic model. MAPref stands for the children adapted model. MAPspk stands for the speaker dependent models.*

For evaluation purposes, an annotation of the corpus was made to know which utterances contain pronunciation errors. This manual annotation has been carried out by a group of independent labelers. Every phoneme in every word of the corpus has been labeled as correct or incorrect by three different judges, and has been finally labeled as correct or incorrect considering the majority of votes of the three judges. In this annotation, a 17.41% of the phonemes in the impaired speech corpus have been labeled as incorrectly uttered.

4.2. Results in ASR

A set of experiments in ASR were carried out over this database with the same specifications of the ASR system and HMM acoustic modeling that are used in the speech therapy application as shown in Section 3.3. Results are shown in Table 1.

The first acoustic model was obtained via the Maximum Likelihood (ML) algorithm from the utterances contained in the databases SpeechDat-Car and Albayzin containing adult Spanish speech and it is the same model used in the Vocaliza application. The baseline results for the 14 speakers give an average Word Error Rate (WER) of 52.22%. Variability in the results among speakers is high, as the speaker with the worst results obtains a 89.47% in WER, while the speaker with the best results obtains a 13.16% in WER. This variability is related to the deep variability in their kinds and degrees of impairment.

A second acoustic model adapted to children speech was trained over the non-impaired speech database via MAP adaptation [5] and tests were carried out over this model with the impaired speech database. The results obtained show a decrease in WER to an average 31.45% among the speakers. Although not initially in the application, this model could be easily included in the application as a way to reduce recognition errors without the use of speaker adaptation when enough data is not available.

Finally, a set of experiments with the MAP algorithm for speaker adaptation implemented in the application was carried out. In this case, a strategy of leave-one-out was taken; this is, every one of the four sessions of every speaker was used for testing a model trained with the speech in the three remaining sessions of the speaker. Average WER among all the speakers in this experiment drops to 16.07%.

4.3. Accuracy detection results

At this moment, a evaluation of the ability of the ASR system has been made. A reduction of the WER has been proved by the use of speaker dependent models. Regarding the speech therapy application, the improvement in the performance of the ASR system avoids the user getting frustrated of been rejected even when he utters perfectly the word. But the objective of the application is to help people with disabilities to correct their pronunciation errors. Because of this, a WER of 0% would be of no use if the speaker is making a great number of mistakes, as the system would not be helping him to correct them.

# of Incorrect phonemes	1	2	3
Words labeled as incorrect	47.72%	21.87%	10.59%
Accuracy (ML)	69.96%	63.66%	56.64 %
Accuracy (MAPref)	71.31%	77.91%	75.47%
Accuracy (MAPspk)	65.41%	87.66%	86.94%

Table 2: *Correct/incorrect detection accuracy for different acoustic models. ML stands for the adult acoustic model. MAPref stands for the children adapted model. MAPspk stands for the speaker dependent models.*

Thus, a new measure is needed to know the real performance of the speech therapy system. This measure has to tell the accuracy of the system to discriminate correct pronunciations from incorrect pronunciations. For this purpose a traditional way to measure accuracy in acceptation/rejection systems was taken. In this case, the accuracy (Acc) was considered as the number of mispronunciations not recognized (TN) (this is, the system correctly does not accept them as pronunciations of the given word) plus the number of correct utterances recognized (TP) (the system accepts them as correct) divided by the total number of utterances (U).

$$Acc = \frac{TP + TN}{U} \tag{1}$$

Although it is clear to define when a phoneme is mispronounced, it is not so clear the definition of a mispronounced word, specially in the case of ASR. Considering a word as incorrectly uttered when at least one phoneme is mispronounced would make 47.72% of the words in the speech impaired database fit that definition. If at least two mispronounced phonemes are required to consider a whole word as incorrect, that would lead to a 21.87% of incorrect words. Furthermore, taking at least three mispronounced phonemes to consider a word as incorrect would mean a 10.59% of mispronounced words. Going further in this classification (at least four incorrect phonemes in a incorrect word) is a not significant case as less than 6% of the words would fit as incorrect and there are some words with only three phonemes who would never be mispronounced.

Table 2 shows the number of incorrect words according to every possible definition of what a mispronounced word is. Also, it shows the results in the accuracy of the ASR system to recognize or not the uttered words according to their condition as correctly or incorrectly pronounced words. Results show that considering that a mispronounced word contains at least two incorrect phonemes the accuracy rises to 87% when using speaker dependent models.

4.4. Discussion

Relevant discussion can be made out of the results achieved. The most relevant is the fact of how accuracy rises from 60% to 76% just by changing the adult speaker independent model to a children speaker independent model; and then to a 87% by using speaker dependent models. This shows that the use of speaker dependent models really is important for a better operation of the system. This is due to the fact that the speaker adaptation process obtains a speaker dependent model that eliminates the errors due to the inter speaker variability in the recognition phase, but speaker adaptation by itself is unable to learn the strong mispronunciations of the speaker, so those errors stay in the system.

However, these results achieved for the cases in which a minimum of two or three incorrect phonemes are set for a word to be considered mispronounced do not follow the same trend for the case of a minimum of one incorrect phoneme. This is due to the chosen vocabulary, the 57 words of the RFI, that does not provide words with a high confusability. That is to say, every pair of words in the vocabulary differs in at least two phonemes (and usually in more than three phonemes). Because of this, when only a phoneme is mispronounced in a given word the system keeps recognizing that word as it is still the closest word in the vocabulary.

5. Conclusions

As a result of this work, a totally functional application which aims to help the work of the speech therapists in three levels of the language (phonological, semantic and syntactic) has been developed. The software is ready to be distributed at this moment, and is free to use for every speech therapist who may require it.

All the requirements set at the beginning of the work have been completely fulfilled. Furthermore, AAC methods embedded in Vocaliza provide added value to the application, making possible to use it as a educational software, not only for people with communications disorders but for people with cognitive disorders or even for young people without disorders.

Also, it has been shown how the HLT included in the application really can help the user to effectively detect pronunciation errors and correct them. Particularly, the use of speaker adaptation rises the accuracy rate of the system according to a manual labeling from 60% to 87%.

6. Acknowledgements

The authors want to thank Prof. Richard C. Rose from McGill University in Montréal (Canada) for his fruitful discussion and ideas that have lead to some of the results in this work.

7. References

[1] Protopapas A. Öster A-M, House D. and Hatzis A., “Presentation of a new EU project for speech therapy: Olp (ortho-logo-paedia)”, TMH-QPSR vol. 44, Fonetik, 2002.

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

[3] García Gómez et al., “Isaeus speech training for deaf and hearing-impaired people,” Tech. Rep., 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] Monfort M., Juárez-Sánchez A., “Registro Fonológico Inducido (Trajetas Gráficas)”, Ed. Cepe, Madrid, 1989.