VOCALIZA: AN APPLICATION FOR COMPUTER-AIDED SPEECH THERAPY IN SPANISH LANGUAGE

Carlos Vaquero, Oscar Saz, Eduardo Lleida

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

José Manuel Marcos, César Canalís

Colegio Público de Educación
Especial Alborada, Zaragoza
cpeealborada@gmail.com

ABSTRACT

This paper introduces Vocaliza, a software application for computer-aided speech therapy in Spanish language. 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 language pathologies. The application works at three levels of language: phonological, semantic and syntactic. For this purpose, the application makes use of four speech technologies: Automatic Speech Recognition, speech synthesis, speaker adaptation and utterance verification. Vocaliza is a flexible and modular application, which will be growing in functionality during the next years. The evaluation given by a group of speech therapists to the work of Vocaliza is encouraging, as long as they consider the application useful and easy to use, which were the main objectives in the development of the application.

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 pathologies. 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 speech technology and speech therapy such as Orto Logo-Paedia [1], SPECO [2], ISAEUS [3] and HARP [4] have been developed, some of them resulting in the development on a software for speech therapy at the end of the research process. However, there are no versions available of these softwares 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. Moreover, in these projects, the work of speech therapy is carried out only at phonological level, when most of speech therapists would be interested in the training of other characteristics of the spoken language.

This work has been supported by the national project TIN 2005-08660-C04-01

This paper 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 way of operation of the software while section 4 shows the application graphical user interface. Section 5 explains how speech technologies are included in the software for speech therapy. Then, in section 6, the evaluation of the software given by a group of speech therapists is explained, to finally extract the conclusions to this work in Section 7.

2. OBJECTIVES AND REQUIREMENTS

The objective of this work was the development of a free-to-distribute software application for speech therapy in Spanish language. For this purpose, suggestions and requirements of specialists in speech therapy were listened prior to the start of the work. The requirements set for the application can be separated from four points of view:

- In terms of the linguistic level that the application should train, there was a special interest in the fact that the application helped for the training in several levels of language: from purely phonological level to syntactic and semantic levels.
- Regarding application usability, the application should provide enough flexibility for speech therapists to work on a wide range of pathologies, while methods used to treat these pathologies should be amusing to attract users (mainly children) with pathological speech.
- The application should have a modular way of dealing with the users, this is, the information from every user should be easy to move between different computers in which the software is installed, without any loss of the user information.
- The application should be easy to use, as speech therapists and patients of speech pathology 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.

For this kind of work, the collaboration of experts in speech therapy and pedagogy is strongly necessary. In our case, we have counted with the assistance of the staff of the Public School for Special Education "Alborada", located in Zaragoza, Spain. Their knowledge in different fields of work with disabled children was essential for reaching the objectives of this work.

3. WORKING WITH VOCALIZA

Vocaliza operation can be summed up in a block diagram as shown in Figure 1. The application is divided into three main blocks with different functionalities, in which the exchange of information between blocks is represented by different arrows. The upper block represents the user of the application, who will manage and use all the options in the application. Each arrow stands for a different type of information:

- **Solid black arrow** represents the configuration information, as Vocaliza enables users to set up different aspects of the application in order to adapt it to every speech pathology.
- **Solid grey arrow** is the audio signal. The users of Vocaliza must talk to the application to train their communication skills, so that Vocaliza will analyze and evaluate the evolution in the speech of the user.
- **Dotted grey arrow** represents the information of the acoustic models. Vocaliza uses Automatic Speech Recognition (ASR) to enable users to train their speech, and it provides flexibility allowing the user to select which model the application will use, making possible to work with both speaker dependent and speaker independent systems.
- **Dotted black arrow** stands for word, riddle and sentence information. Vocaliza uses isolated words to train phonological language aspects, riddles to train the semantic level of language and sentences to train the syntactic level. Every word, riddle and sentence belonging to the application must have related information such as an image or text.

3.1. Games: Speech training

The games block provides different speech training methods, enabling to train every one of the three language levels that were set in the requirements of the software. The best way to train speech is speaking, so the speech training in Vocaliza is based on making the user speak. It is not difficult for an adult person to understand that speaking is a great way to improve his/her communication skills, but a child could find any speech therapy hard, tiring and boring. This is the reason for Vocaliza to use games to train user communication skills, making the application interesting to children with pathological speech.

Vocaliza features four different games to train user speech. Three games are designed to train specific levels of language, whereas the fourth game provides additional functionality. All the games follow the same scheme, the application shows some visual information to the user while playing an audio recording or synthesized speech. The user must utter the word or sentence related to the audio-visual information in order to complete the game successfully. Therefore, every game needs ASR to work properly.

The four games included in Vocaliza are:

- **Pronunciation game**: Pronunciation game is designed to train phonological language level, working with isolated words. To complete this game, the user must utter correctly the word shown on the computer screen. Then, the application will evaluate user pronunciation, displaying a score on the screen. This score will depend mainly on the improvement shown by the user in his/her pronunciation.
- **Riddle game**: Riddle game is designed to train semantic aspect of language. This game works with riddles, which are composed by a question and three possible answers. Every answer is an isolated word. The user must choose the right answer and pronounce it correctly to complete the game successfully.
- **Sentence game**: Sentence game is designed to train syntactic level of language. It works with sentences and the user must pronounce all the words in the sequence on the correct order to complete the game.
- **Evocation game**: Evocation game is designed to enable users to pronounce freely every word they might want to train. Unlike all the other games in Vocaliza, evocation game does not provide any visual or audio information to the user but a visual feedback of the word uttered by the user. As the pronunciation game does, evocation game works with isolated words and trains mainly phonological level of language.

As shown in Figure 1, the users of Vocaliza can configure the games. The application allows users to decide if they want to listen to the audio information or if the application must display any kind of text related to every word, riddle or sentence that the users may face. This enables speech therapist to adapt Vocaliza to every user requirements, depending on their communication and cognitive skills

In addition, games take information about the words, riddles and sentences that the user wants to train, as well as the acoustic model for that user (speaker dependent or speaker independent), providing flexibility to treat a wide range of speech pathologies.

3.2. Training: Speaker Adaptation

Speaker adaptation allows users to train speaker dependent acoustic models. All the games of the application

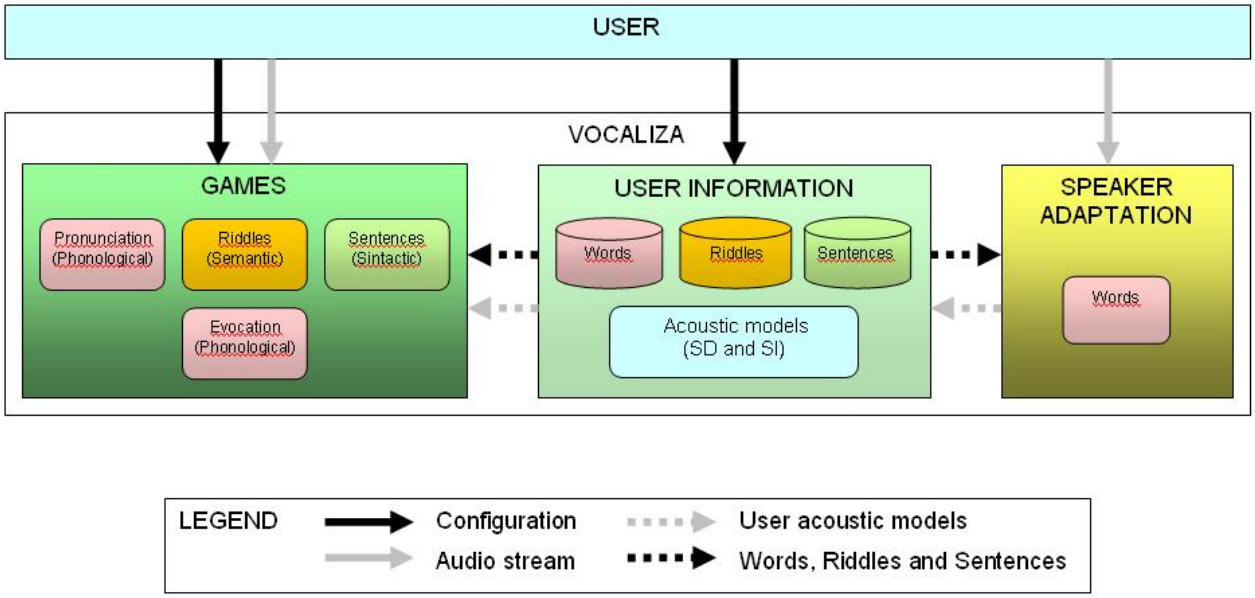

Figure 1. *Vocaliza: Block Diagram.*

are based on ASR systems, and the use of speaker adaptation is necessary to improve the performance of the ASR system when the user suffers a severe pathology. In the pronunciation game, the speaker dependent model is also necessary to give out the evaluation to the speaker speech, enabling the user to track his/her evolution and communication skills improvement.

Speaker adaptation in Vocaliza works with isolated words. The user selects which words to pronounce and how many times to utter each word, and starts recording his/her utterances (Vocaliza provides recording functionality). With this recorded data, the application will train an acoustic model adapted to the user.

As Figure 1 shows, speaker adaptation block take words stored in user information block and speech utterances of these words from the user, to train an adapted acoustic model which will be stored in the user information block.

3.3. User Information

The user information block stores all information regarding user configuration, including all words, riddles and sentences the user may face, the default configuration for the games, and every acoustic model available to the user.

Every person training his/her speech with Vocaliza should have his/her own user inside the application, in order to keep and store all the setup information related to his/her speech training. This provides great flexibility and modularity, allowing the user to add new words, riddles and sentences to train himself, as well as to select the acoustic model that the ASR system will use, being able to choose between speaker dependent or speaker independent acoustic models.

Figure 1 shows that this block is fully managed by the user, and it provides all the user information and configuration for the games. In addition, it brings word information to speaker adaptation block and stores every speaker dependent acoustic model.

All information associated with a user in Vocaliza is stored in different files which are kept in the same folder in the computer. Since users might want to train their speech in different places, and thus, in different computers, user information has been organized so that carrying it is as easy as copy the folder that contains all the information of the user from one computer to another.

4. GRAPHICAL USER INTERFACE

The Graphical User Interface (GUI) in Vocaliza was designed keeping in mind that any user with speech pathologies should be able to use the application without any help, even if the user is not used to handle computer applications. Moreover, Vocaliza GUI should provide an easy interface to manage all setup options, since speech therapists may not be used to work with computers.

Therefore, Vocaliza main window layout was designed as shown in Figure 2. This window has two important and distinct regions: the main menu, containing all configuration options and managing functionality, and the image panel, which provides easy access to every game in Vocaliza. This layout enables children with speech pathologies to play games easily, without adult supervision, using only the image panel, whereas speech therapist can access all configuration options fast and without trouble using the main menu. If the user selects a game, simply clicking on one of the four game images placed at the bottom half of the main window, the corresponding game window will open, showing all images and text related to the selected game.

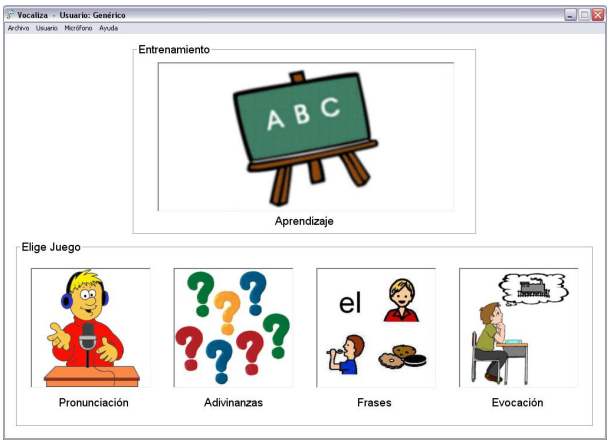

Figure 2. *Vocaliza: Main Window.*

Figure 3. *Vocaliza: Riddle Game.*

Games GUI is designed in order to avoid any user interaction different from speaking at the microphone or listening to the application speech, so user will not need to know how to use a keyboard or a mouse to play games. Only for closing the game window the use of another input mode different from speech is required. Figure 3 shows a screenshot of the riddle game.

Speaker adaptation process is designed to be as easy as possible too, but some knowledge about managing windows applications may be required, so this process is intended to require adult supervision. Speaker adaptation process can be started by clicking on the blackboard image at the top half of the main window, and it takes three steps:

- First of all, the user will select the word to pronounce and will record an utterance of the selected word, using the training window, as Figure 4 shows.
- Then, the user will check if the utterance was recorded with enough quality to use it to estimate an adapted model. For this purpose, the signal window enables the user to see and listen the recorded signal. The signal window is shown in Figure 5. These first two steps may be repeated several times, to get as many utterances of speech from the user as desired.

Figure 4. *Vocaliza: Training Window.*

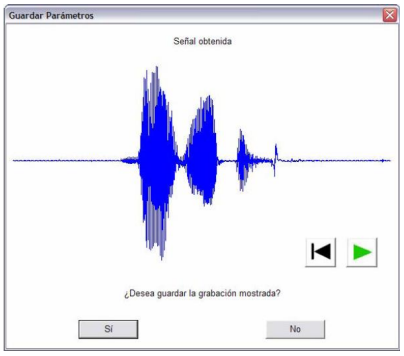

Figure 5. *Vocaliza: Signal Window.*

- Finally, the user will select which utterances will be used to estimate the adapted model, using the utterance selection window, as shown in Figure 6, and the speaker adaptation process will start.

User data and user configuration can be fully managed through user window, which is shown in Figure 7. The access to this window is done through the main menu and allows the user to add or remove words, riddles or sentences, as well as to configure all aspects of games behaviour and to select the acoustic model that ASR will use.

5. SPEECH TECHNOLOGIES IN VOCALIZA

Most of Vocaliza functionality is provided by speech technologies. Vocaliza games use an ASR system to decide if the user has completed the game successfully, speech synthesis to show how a word must be pronounced, speaker adaptation to estimate acoustic models adapted to the user, and utterance verification to evaluate user pronunciation.

5.1. Automatic Speech Recognition

Automatic Speech Recognition is the core in the Vocaliza application. Every game needs ASR to decode the user utterance, 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.

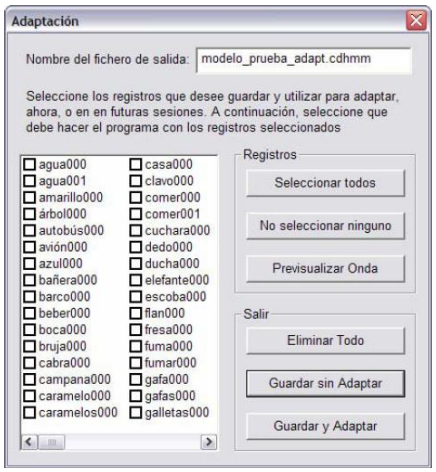

Figure 6. *Vocaliza: Utterance Selection Window.*

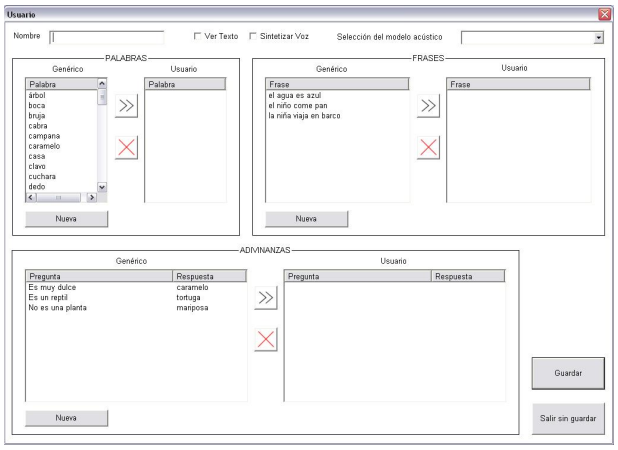

Figure 7. *Vocaliza: User Window.*

Speech signals are adquired 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 Coefficient), consisting on 12 static parameters, 12 delta parameters, 12 delta-delta parameters and the delta-log-energy. The units used by the system are 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 modelled with 1 state per model and a 16-Gaussian mixture for every state.

5.2. Speech synthesis

Speech synthesis provides a method to show the user how a word or sentence should be pronounced. It is not the only method available in Vocaliza but is the easiest method to use: as soon as a new word, sentence or riddle is added to the application, Vocaliza 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 utterances, which the application will use instead of speech synthesis, in order to show different utterances depending on user age, speech pathology and other requirements of the user. For instance, slow utterances could be shown to train very young children speech.

5.3. Speaker Adaptation

Speaker adaptation enables Vocaliza 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, not like other estimation methods such as Maximum Likelihood (ML) training. Not needing a great amount of data from the user is a very interesting feature since Vocaliza 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 every user speech, and users with pathological speech will find very hard and tiring to record a great amount of speech. Moreover, accuracy in speaker dependent ASR based on MAP estimation methods tends to be equal to accuracy in speaker dependent ASR based on ML estimation methods when the number of utterances is high, so the use of MAP will not involve any loss in the ASR performance.

At first, there is only one acoustic model for all users in Vocaliza, which MAP will use as starting point to estimate all adapted acoustic models. As every user records his/her utterances and launches speaker adaptation, he/she will get better adapted acoustic models.

5.4. Utterance Verification

Utterance Verification (UV) is a technique embedded in Vocaliza to provide a mechanism to evaluate how the speech of the user improves. This technique tries to quantify how much has the user improve his/her communication skills.

Classically, UV has been proposed as a mechanism to verify accuracy of recognition hypotheses, which assigns measures of confidence to each hypothesized segment obtained through Viterbi segmentation of the utterance. Typically, a measure of confidence is assigned to every recognized word, and each hypothesized word is accepted or rejected depending on its corresponding measure of confidence.

Vocaliza uses a Likelihood Ratio (LR) based UV [6] procedure to assign a measure of confidence to each hy-

pothesized 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, in addition, roughly quantifies user speech improvement.

In order to obtain a confidence measure to quantify user speech improvement, Vocaliza 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 pathological speech, as alternate hypothesis. Assuming that the only significant variation that user speech will undergo during his/her treatment will be an improvement regarding his/her pathological speech, the measure of confidence obtained for every word in every utterance will increase as user improves his/her communication skill. Therefore, this measure of confidence involves a relative evaluation method to quantify user speech improvement, which will need a initial confidence measure value to work correctly.

6. EVALUATION OF VOCALIZA

To evaluate the usefulness of Vocaliza as well as all the features that the application was intended to provide (usability, modularity and flexibility), the first released Vocaliza version was handed to teachers and speech therapists in "Alborada" School. In addition to the application, two questionnaires were given to the staff of the School. The first one was aimed to the teachers and speech therapists in order to retrieve their opinion about the usefulness, modularity and flexibility of the application, whereas the second one aimed to collect the opinion of children and pupils from "Alborada" School about usability and attractiveness of Vocaliza.

All the questionnaires retrieved show the great interest that the staff of "Alborada" School has for Vocaliza. Both teachers and pupils consider Vocaliza a useful and entertaining application. Teachers from the School not only think Vocaliza is useful as a application for computer-aided speech therapy but as a educational application as well. They find the riddles game very useful for the training of cognitive skills of children with speech or without speech pathologies. Furthermore, the way of displaying the images in the sentences game is very similar to some systems of augmentative and alternative communication (AAC), which makes Vocaliza also very interesting as AAC system.

On the other hand, pupils from "Alborada" find the application amusing, but they admit having troubles to use the application on their own. However it is important to remark that those pupils are not used to work with computers at all and they also suffer cognitive disorders associated to their speech pathology.

7. CONCLUSIONS

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

The evaluations by a group of speech therapists shows that Vocaliza is a useful tool for the training of children with speech disorders at several levels of the language. The speech therapists also evaluate positively the easiness of use of the software, altogether with the usefulness of the speech technologies implemented in Vocaliza for their work. In this way, all the requirements set at the beginning of the work have been completely fulfilled.

All these reviews and evaluations are very encouraging for us to keep working in this direction as we plan to keep improving the functionality of Vocaliza.

8. BIBLIOGRAPHY

[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," Tech. Rep., 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.