Speech Technology Applied to Children with Speech Disorders

W. Ricardo Rodríguez[1], Carlos Vaquero[2], Oscar Saz[2], Eduardo Lleida[2]

[1] University of Zaragoza , Communication Technology Group (GTC), I3A, Zaragoza, Spain
[2] University of Zaragoza , Communication Technology Group (GTC), I3A, Zaragoza, Spain

Abstract— **This paper introduces an informatic applications for speech therapy in Spanish language based on the use of Speech Technology. The objective of this work is to help children with different speech impairments to improve their communication skills. Speech technology provides methods which can help children who suffer from speech disorders to develop pre-Language and Language. For pre-Language development the informatic application works on speech activity detection and intensity control, and for Language the application works on three levels of language: phonological, semantic and syntactic. This application is designed to enable those suffer from speech disorders to train their communication capabilities in an easy and entertaining way. This tool is available for free distribution.**

Keywords— **Pre-Language, Language, speech processing, speech therapy.**

I. Introduction

Over the past decade there has been a growing interest in the use of Speech Technology (ST) for teaching of pronunciation, speech therapy and researches. ST covers different work fields and areas of application. The most widely spread tendencies are, among others: Speech processing, natural speech processing, dialogue systems and Linguistics. This technology, which are becoming more and more common to us and help us make our life simpler, could be a great help for people with speech disorders. The aim of this study is helping the daily work of speech therapists applying part of ST to allow people with speech disorders to communicate better.

In some cases, language is not developed adequately during the first year of life; in others, due to different disorders, the individual has difficulties with pronunciation, articulation or word creation, which inhibits communication. The development of informatic applications based on such technology allows individuals with speech disorders better communication and interaction with their environment, even with computers.

In section II, this article shows an informatic application which uses graphic animation and speech signal processing to develop or improve pre-Language in children with speech disorders. Section III explains how ST may be used to enable the development of Language.

II. Pre - language

The first signs of pre-Language include crying, speech activity detection, intensity and intonation control, generation of vocal sounds and others, which take place during the first year of life. After this stage, the development of Language continues until the age of five. During which period the child acquires the essential part of language which will allow it to fully master it in time [1]. This situation occurs in healthy children, but unfortunately, there is a large number of cases of children with speech disorders, who do not even develop adequate pre-Language. To improve pre-language, an informatic applications has been developed which works on aspects like speech activity detection and intensity control. By processing the speech signal, parameters are obtained to control graphic animations (in a playful way) and create awareness in children of their own speech. The detection, processing of speech signal and corresponding animations are implemented in C++.

A. Speech activity detection

The ST used is a Voice Activity Detector (VAD), the primary function of a VAD is to provide an indication of the presence of speech in order to facilitate speech processing as well as possibly providing delimiters for the beginning and end of a speech segment.

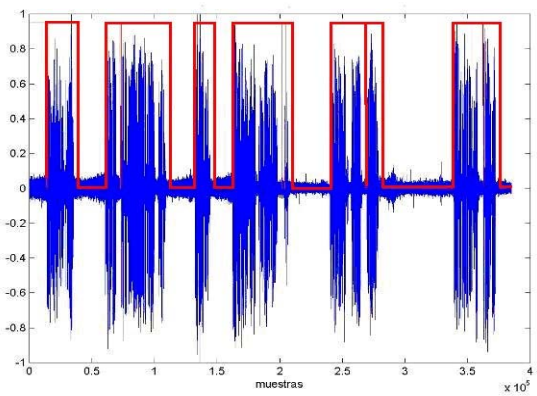

Fig. 1 Speech signal and VAD decision.

Fig 1. shows the shape of the speech signal (blue) and decision of the VAD for a voice segment (red line).

The objective is the creation of graphic animations according to the presence of speech Voice segments (values red line = 1) are used to control variables within C++ routines, developed with Allegro graphic libraries, since these are designed for games.

One of the computer applications shows a black background (Fig 2. Left) and lines and triangles in different colors and positions begin to appear when voice is detected. This figures moves randomly around the window drawing the path they followed (Fig 2. Right). The child will become aware of the possibility of drawing by use of voice.

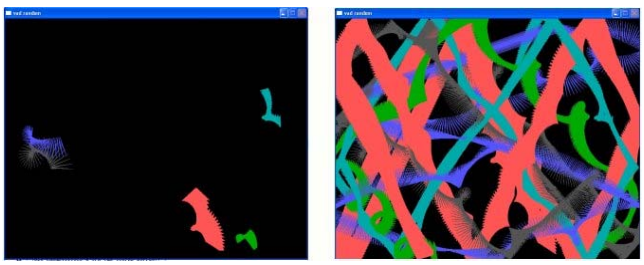

Fig. 2 Drawing by use of voice.

Another computer application uses the voice detected to move a car. At the beginning the car appears the left side, when voice is detected it moves to right side (Fig 3). Similar application is used in intensity control.

B. Intensity control

Controlling intensity and breath is important to communicate. In order to develop the child's ability to control intensity and breath, the technique used is the energy estimation of speech signal.

Fig. 3 Car moved by voice .

In particular, the amplitude of unvoiced segments is generally much lower than the amplitude of voiced segments like in Fig 4 (up). The short time energy of the speech signal provides a convenient representation that reflects these amplitude variations [2] (Fig 4 down).

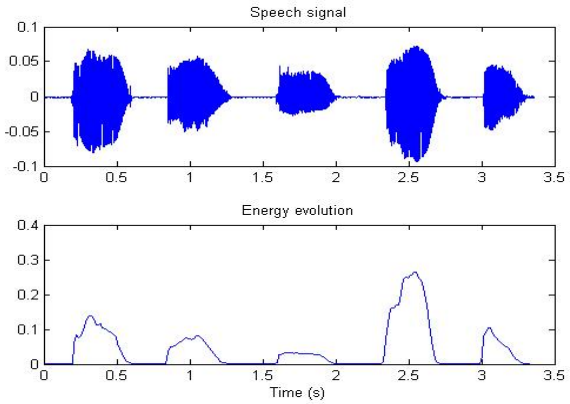

Fig. 4 Speech signal and energy evolution

The value of the energy is used to control variables in graphical routines so that an increment of energy, becomes an increment velocity in the same car of section *A*; So, the child can associate the intensity of voice with the control of the speed of the car.

In order to induce the child to control (modulate) voice intensity, in another informatic application the child has to fly a cartoon (Dragon) Through a maze. In this case an increment of energy, becomes an increment of the *y-axis* position Dragon and vice versa, a decrement of energy of the speech signal becomes a decrement of the *y* position dragon. Speech activity detection becomes movement of the *x-axis* position Dragon. Fig 5. shows the Dragon through the maze.

Fig. 5 Dragon through the maze.

III. Language

ST may be very useful as well to contribute to the development of language in individuals who suffer from some sort of speech disorder or simply in children who are at the stage of developing language. Furthermore, ST can support the daily work of the speech therapists who train the linguistic skills of Spanish speakers with different language pathologies. Below, *Vocaliza* [3], an application for computer aided speech therapy in Spanish language is introduced. It is based on ST, and provides a user interface especially designed to be attractive to even the youngest users, as shown in Fig 6. Vocaliza works at three levels of language: phonological, semantic and syntactic. Each level is trained by a different method which is shown as a game, in order to attract young users.

The *Phonological* level is trained encouraging the user to pronounce 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's speech pathology. The application evaluates every user utterance and displays a mark with a cartoon on the screen, that end user will be able to understand easily.

The *Syntactic* level is trained by 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 mark, showing the user their improvement.

The *Semantic* level is trained by means of a set of riddles, previously defined by a speech therapist or pedagogist. The application asks the user a question and gives three possible answers. The user must utter the correct answer to go on with the next riddle. The application will again show a mark depending on user utterance.

All games are based on automatic speech recognition, (ASR) which will decide if the word or sentence uttered by the user is that which the application was expecting. Most of *Vocaliza* functionalities are provided by ST. Vocaliza makes use of ASR 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.

A. Automatic Speech Recognition

Automatic Speech Recognition is the core of the *Vocaliza* application. Every game needs 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. 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 Coefficients), consisting on 12 static parameters, 12 delta parameters, 12 delta-delta parameters and the delta-log-energy.

The acoustic model is composed of a set of 822 context-dependent units plus a silence model and an inter-word 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. The ASR system embedded in Vocaliza uses a utterance verification procedure to decide if the user has pronounced the requested word or if there is a phoneme sequence with more probability.

B. Speech Synthesis

Speech Synthesis provides a way 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 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, riddle, and sentence 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 speech of very young children.

C. Speaker Adaptation

Speaker adaptation enables *Vocaliza* to estimate speaker dependent acoustic models adapted to each user. Vocaliza uses Maximum A Posteriori (MAP) estimation [4] 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 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 speech of every user, and users with pathological speech will find very hard and tiring to record a great amount of speech utterances. 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.

D. Utterance Verification

Utterance Verification (UV) is a technique embedded in the application to provide a mechanism to evaluate the improvement of user communication skills. 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 [5] 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, 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 in the user speech 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 the user improves his/her communication skill.

IV. Results

As a result, there are two Computer aided applications, one of pre-Language *(Speech Activity Detection and intensity control)* and one for Language (Vocaliza), and they are both in use at a special education school in Zaragoza, which works with different cartoon programs. They have been very successful among the children and speech therapists, even in cases where speech therapy initially did not work, but *Speech Activity Detection* allows early sensory stimulation and attention to be caught. The evaluations by the group of speech therapists shows that these informatic applications are a useful tool for the training of children with speech disorders at several levels of the language. The therapists also evaluate positively the easiness of use of the applications. The results are very encouraging to keep working in this direction as it is planned improve the functionality and robustness of informatic applications.

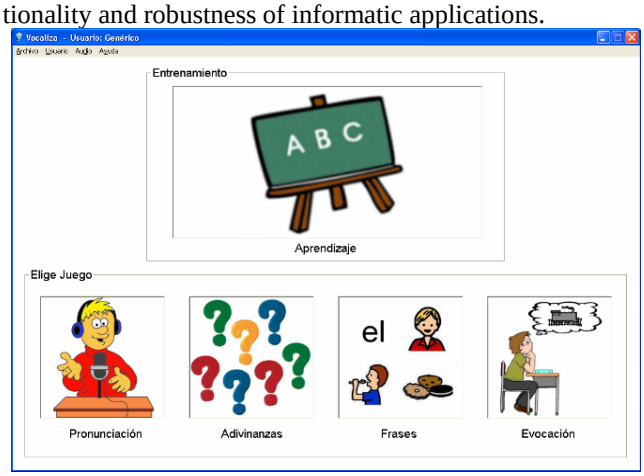

Fig. 6 Vocaliza main window.

V. Conclusions

ST is an essential tool to directly and easily support the development and training of children who suffer speech impairments. The adequate development of Language and pre-Language improves the quality of life of individuals with speech disorders and enable them to use computers by means of multimedia applications.

Acknowledgment

This work has been supported by MEC of Spanish government through National Project TIN 2005-08660-C04-01 and Santander Bank Scholarships - University of Zaragoza.

References

1. Puyuelo M (1996) Evaluación del Lenguaje. Masson, España
2. Rabiner L, Schafer R (1978) Digital Processing of Speech Signals. Prentice – Hall pp. 116-120.
3. Vaquero C, Saz O, Lleida E, Rodríguez W (2008) E-inclusion Technologies for the Speech Handicapped, ICASSP Las Vegas, USA ‘‘in press’’
4. Gauvain J, Lee C (1994) Maximum a Posteriori estimation for multivariate Gaussian mixture observations of Markov chains, IEEE Transactions on Speech and Audio Processing, vol. 2, pp 291-298
5. Lleida E, Rose R (2000) Utterance verification speech recognition, IEEE Transactions on Speech and Audio Processing, vol. 8, no. 2, pp. 126–139

Author: William Ricardo Rodríguez Dueñas.
Institute: I3A, GTC, Universidad de Zaragoza.
Street: Maria de Luna, 1.
City: Zaragoza.
Country: España.
Email: wricardo@unizar.es