Available online at www.sciencedirect.com

ScienceDirect

Speech Communication xxx (2011) xxx–xxx

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A prelingual tool for the education of altered voices

William R. Rodríguez*, Oscar Saz, Eduardo Lleida

Communications Technology Group (GTC), Aragón Institute for Engineering Research (I3A), University of Zaragoza, María de Luna 1. 50018, Zaragoza, Spain

Abstract

This paper addresses the problem of Computer-Aided Voice Therapy for altered voices. The proposal of the work is to develop a set of free activities called *PreLingua* for providing interactive voice therapy to a population of individuals with voice disorders. The interactive tools are designed to train voice skills like: voice production, intensity, blow, vocal onset, phonation time, tone, and vocalic articulation for Spanish language. The development of these interactive tools along with the underlying speech technologies that support them requires the existence of speech processing, whose algorithms must be robust with respect to the sources of speech variability that are characteristic of this population of speakers. One of the main problem addressed is how to estimate reliably formant frequencies in high-pitched speech (typical in children and women) and how to normalize these estimations independently of the characteristics of the speakers. Linear prediction coding, homomorphic analysis and modeling of the vocal tract are the core of the speech processing techniques used to allow such normalization through vocal tract length. This paper also presents the result of an experimental study where *PreLingua* was applied in a population with voice disorders and pathologies in special education centers in Spain and Colombia. Promising results were obtained in this preliminary study after 12 weeks of therapy, as it showed improvements in the voice capabilities of a remarkable number of users and the ability of the tool to educate impaired users with voice alterations. This improvement was assessed by the evaluation of the educators before and after the study and also by the performance of the subjects in the activities of *PreLingua*. The results were very encouraging to keep working in this direction, with the overall aim of providing further functionalities and robustness to the system.

Keywords: Altered voice; Voice therapy; Speech processing; Formant normalization; Vocal tract length estimation

Contents

* Corresponding author. Tel.: +34 976 762705; fax: +34 976 762111.
E-mail addresses: wricardo@unizar.es (W.R. Rodríguez), oskarsaz@unizar.es (O. Saz), lleida@unizar.es (E. Lleida).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.05.006

1. Introduction

Voice therapy involves extended interaction between an individual user and a skilled therapist in order to train or to educate the user's altered voice. In general, altered voice means any acoustical disturbance which affects the voice quality, either the intensity, fundamental frequency, formants, duration, or various combinations of them (Arias and Estape, 2005; Aronso, 1993). This alteration may be caused by structural anomalies, such as weakness of laryngeal structures or muscles; by pathological conditions including vocal nodules, polyps, vocal fold thickening, by vocal abuse or vocal misuse, and by speech disabilities in cases of handicapped individuals (Kenneth, 1966; Kornilov, 2004). Initially the patient must be examined by a laryngologist and a speech therapist, the consultation between them may result in the recommendation of voice therapy and may include the use of computer-aided tools for this purpose.

The development of these tools for voice therapy has been a major issue since the 1990's. Speech Technologies have been increasing their robustness to different ambient conditions and this has allowed the creation of them. Some of the most remarkable efforts in this area comprehend the following applications:

- SpeechViewer III, developed in 1997 by IBM (Speech Viewer), was the most popular commercial software for speech therapy, although currently is not having any more support from its company. It relied strongly in the robustness of signal processing like voice detection or pitch tracking. This application used a very simple speech interface, which allows to train voice activity detection, varying intensity, pitch and vocal onset.
- Vox Games, developed by CTS Informatica in Brazil (VoxGames), is a pack of 25 games specially developed to clinical therapy, with the aim of stimulating the voice and speech modification in children and young adults, for the purpose of promoting a better production and control of various parameters. The modules for the vocal practice are games for intensity, pitch, phonation time, voice activity detection and voice onset.
- Games Program, M5176, by KAY Elemetrics Corp. (Games Program), provides an environment for speech therapy which focuses on the control of intensity and pitch amplitude. Each game can be modified by the clinician to make a task more or less challenging, depending on the user's level of performance.
- Dr. Speech, developed by Tiger DRS (Dr. Speech), is a comprehensive speech/voice assessment and training software system which is intended for use by professionals in voice and speech fields. This tool attempts to produce a reinforcement in changes in pitch, intensity, voice activity detection, vocal onset and vowel tracking. In this case, the feedback to the user is a F1 vs F2 plot representation, which is hardly understandable by non-professional users.

There are some cases in which the resources are not sufficient to provide the acquisition of voice skills in individuals with handicaps due to the expensive price of Computer-Aided tools for voice therapy, the limitations for the training of vocalic articulation and the lack of available tools for a language like Spanish. This situation leads to the use of traditional techniques, for instance, to work vocalic articulation, child and therapist sit in front of a mirror to work praxias (tongue movements) through imitation, or they blow up balloons to work blowing abilities, where the time required to provide this therapy for a single patient can make it impractical when working with a large population.

This paper describes the set of interactive tools that were developed to reduce the time and the level of expertise required from the therapist for providing the interactive component of the therapy. This tool called *PreLingua* covers the first stage of language acquisition (phonatory skills) and include voice activity detection, the control of voice intensity, blow, vocal onset, phonation time, tone, and, vocalic articulation activity in Spanish language.

As the proposed method in *PreLingua* differs completely in their approaches from the traditional methods, it was impossible to compare them objectively. To test the effectiveness of *PreLingua*, a study was conducted in two schools for special education in Spain and Latin America, considering objective measurements from the statistical analysis of the results stored by *PreLingua* and subjective measurements from a therapy evaluation form for each user proposed by the therapist. The results aim to prove that *PreLingua* can actually help patients with speech disorders to improve their voice capabilities, promoting further research in this direction. As *PreLingua* is a free tool available on line, it is possible to benefit a large number of Hispanic-speaking countries from this work.

This paper begins in Section 2 by reviewing the activities developed in *PreLingua*. The speech technologies within *PreLingua* are described in Section 3 and Section 4 addresses the problem of formant estimation in children's speech, and the method to normalize formant frequencies through the vocal tract length. The evaluation of *PreLingua* in real cases is explained in Section 5 and, finally, the results, discussion and conclusions are extracted in Sections 6–8, respectively.

2. Development of voice therapy tools

PreLingua is part of the set of Computer Aided Language Learning (CALL) tools within "Comunica" (Saz et al., 2009). The main goal of "Comunica" is to provide the community of speech therapists with applications that can reduce the time that they need for every one of their patients and students by the automation of many of the activities they carry on every day. "Comunica" intends to be a long-term way of distribution of all the tools for speech and language therapy developed under its framework. All the tools are distributed under a freeware license for the use of all the community of speech therapists and users who could be interested in them (Vocaliza).

All the applications in "Comunica" have been developed for the community of speech therapists in Spain and Latin America by choosing the Spanish language as the working language of operation in these tools.

2.1. PreLingua

Simultaneously to language acquisition, children and adults with severe disabilities need to learn to control their own speech production with voice skills like intensity, blow, vocal onset, phonation time, tone, and vocalization. *PreLingua* is aimed to help to reach this goal and consider these parameters because they are measurable by means of the speech technology.

PreLingua gathers a set of game-like applications that use speech processing to train patients with speech development delays, or special voice needs of handicapped individuals in a special education environment. *PreLingua* is a tool to work with sustained unvoiced and voiced sounds, and its visual interface does not require any previous configuration and is specially designed for children and disabled adults by using visually appealing graphics representing important voice parameters. These voice parameters are obtained from user's acoustical information directly, instead of the use of an automatic speech recognition system like in (Kirschning and Cole, 2007), where they make use of this kind of speech technology in language therapy in children with hearing disabilities.

As shown in Fig. 1, *PreLingua* is designed as a pyramid with five levels according to five rising degrees of difficulty in the activities. The base (level 1) corresponds to *Voice Activity* games, level 2 to the control of *Intensity*, level 3 to *Blow*, *Vocal Onset* and *Phonation Time*, level 4 corresponds to the *Tone* activities and, finally, level 5 to vocalic articulation activities with *Vocalization* and *Articula*. The interfaces for some of these activities can be seen in Fig. 2. *PreLingua* is powered by the Allegro graphic engine (Allegro Graphic Engine).

2.1.1. Voice detection – Level 1

This module provides activities with animated graphics to represent voiced sounds in real time. This feature is especially oriented to children in the very early stages of development who still do not associate their production of sounds to changes in their environment. A Voice Activity Detector (VAD) based on an energy threshold is used by the system (Sakhnov et al., 2009). A binary signal (0:*silence*, 1:*sound*) is the only output of the system. When the voice segment is analyzed and the energy exceeds the established threshold, the sonority of the segment (pitch presence) is studied to ensure that it is a voiced segment when the minimum sonority threshold is reached. When the segment is classified as voiced segment, the system produces a reaction on the screen in which a set of simple shapes and colors will interact around the screen.

In activities like in Fig. 2(a), the system moves the car on the screen only when the voice is detected.

2.1.2. Intensity – Level 2

Once the patient has acquired the ability to distinguish his/her own speech production, it is necessary to learn the control of the volume of this vocal production.

In this case, the value of the intensity in the voice signal is used to modify the position of a certain object or character on the screen. The feedback given by the game is the completion of a goal; for instance, reaching the end of a labyrinth in which a cartoon is controlled up and down accordingly to the intensity of the voice production as in Fig. 2(b). A set of similar games are also provided in *PreLingua* in which the objective of the game is to move

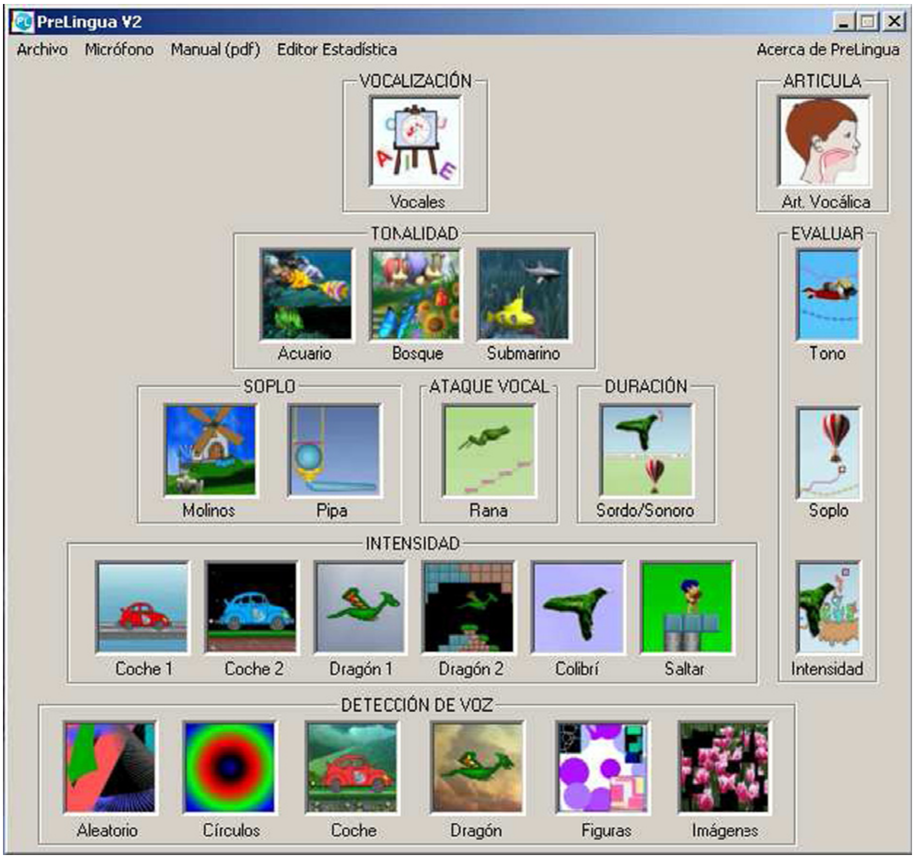


Fig. 1. PreLingua main interface.

the character up and down accordingly to the sound intensity.

2.1.3. Blow, vocal onset and phonation time – Level 3

This level covers three important aspects related to breathing and speech production: *Blow*, *Vocal Onset* and *Phonation Time*.

The *Blow* activities make use of the intensity value estimated in unvoiced (pitchless) segments. The detection of these segments produces an animation on the screen like a blowpipe in Fig. 2(c). In this game the therapist can change the threshold in order to modify the degree of difficulty.

Vocal Onset helps the patient to control the attack of the vocal folds, which is very useful in stuttering cases. In this activity, a character (frog) is controlled by the patient's vocal onset, with the frog jumping in each vocal attack by the user as in Fig. 2(d); in this case, the therapist can set up the space between the bases where the frog jumps.

Phonation Time activity is aimed to help the patient assess their voiced and unvoiced phonation time. Maximum Phonation Time (MPT) and Maximum Exhaling Time (MET) imply abilities in voice production and provides information about the efficiency of the glottal closure (Arias and Estape, 2005). In this activity (shown in Fig. 2(e)) the child is instructed to sustain a voiced sound as long as possible following deep inspiration; this voiced segment produces the flight of a character (bird) and the system indicates the MPT value. After that, the child is instructed to sustain an unvoiced sound as long as possible following deep inspiration, which makes the flight of a second object/character (balloon) and the system shows back the MET value. When the activity is completed the system provides the ratio $\frac{MET}{MPT}$ which can be used to evaluate the glottal closure by the therapist.

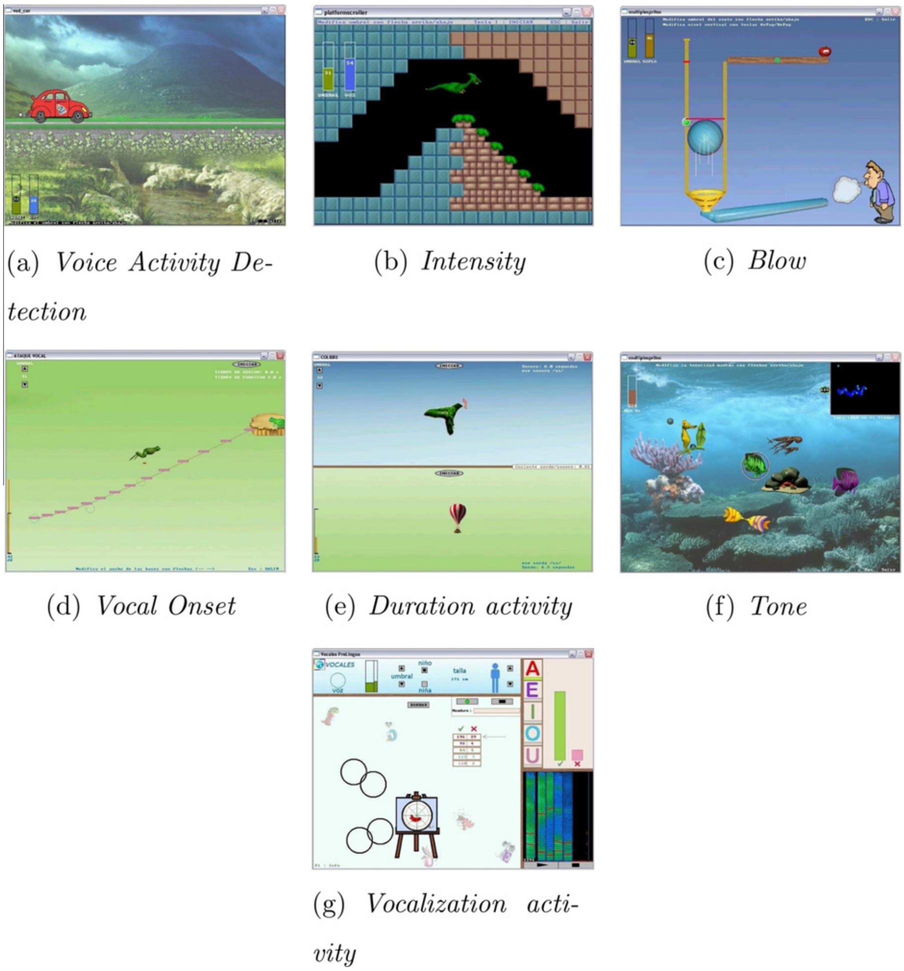
2.1.4. Tone – Level 4

This level is designed to help patients which need to improve pitch control and develop smooth modulation of tone contour. Control of tone is required in a correct speech production and it is extremely needed in some speech features like prosody. In this case, a linear prediction coefficients (LPC) analysis is required to separate the influence of the glottal pulse from the vocal tract. Once the LPC analysis is done (Section 3.1), the system obtains an estimation of the pitch frequency $F0$ from the autocorrelation of the prediction error $d[n]$.

In the case of Fig. 2(f), a fish (center) is moved by the pitch contour and the goal is to follow the other animals on the main stage (aquarium).

2.1.5. Vocalization – Level 5

The transition between phonation and articulation in language acquisition initially occurs with vowels. The set


(a) *Voice Activity Detection*

(b) *Intensity*

(c) *Blow*

(d) *Vocal Onset*

(e) *Duration activity*

(f) *Tone*

(g) *Vocalization activity*

Fig. 2. Some activities in PreLingua.

of vowels for every language is unique, so the strategy in the activities to motivate vocalization has been designed purely for the vowels in Spanish. This language contains five vowels (/*a*/, /*e*/, /*i*/, /*o*/ and /*u*/ in their SAMPA notation) whose representation in the space of the formant frequencies $F1$ and $F2$ is a triangle.

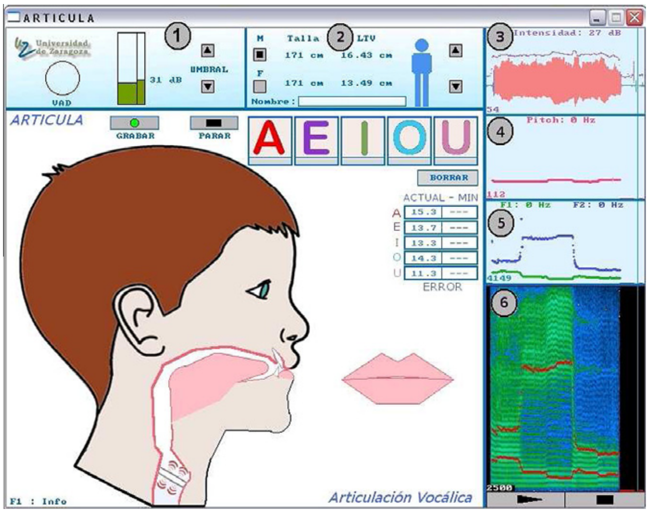
The canonic values of $F1$ and $F2$ for the five Spanish vowels are given by several authors (Martínez-Celdrán, 1989) but these values are purely theoretical and vary largely among different speakers, especially in the case of high pitched voices of women and children. Hence, in this work a formant normalization was applied with the vocal tract length in order to obtain a better estimation for each user.

When the application is running, the therapist selects the gender and height of the user and the system re-adjust the expected formant values for a patient of similar physical characteristics. As showed in Fig. 2(g), the activity allows to train each vowel (vowel /a/ in this case) showing a dartboard, and the position for the darts is the normalized $F1$ against the normalized $F2$ obtained from the user's voice after the normalization process. The goal is to aim and hit inside the desired target region for each vowel. This activity also shows the voice spectrum with the unnormalized formants evolution in real time.

2.2. Articula

A more natural approach to the training of vowels was proposed as a separated activity using a more robust technique for formant estimation that will be fully explained in Section 4. This tool, aimed to show the patient how to position the most basic elements of the vocal tract (tongue, mouth aperture and lips) in the articulation of the different vowels. It is well known that the vocalic sounds are determined primarily by tongue position, the degree of constriction and the lip shape. These characteristics can be correlated to the acoustic features of the vowels which can be identified by two lowest formants, $F1$ and $F2$.

Articula, whose interface can be seen in Fig. 3, uses the two lowest normalized formants F_{1N} and F_{2N}, in order to animate a boy or girl avatar. The extraction of the normalized formants will be explained in Section 4.3 and allows to work with values which are less sensitive to inter-speaker variability.



31 dB

M Talla LTV
171 cm 16.43 cm
F
171 cm 13.49 cm
Nombre :

Intensidad: 27 dB
54
Pitch: 0 Hz
112
F1: 0 Hz F2: 0 Hz
4149
2500

	ACTUAL	MIN
A	15.3	---
E	13.7	---
I	13.3	---
O	14.3	---
U	11.3	---

Fig. 3. Articula. 1 – Threshold voice, 2 – Gender and height selection, 3 – Speech signal and Intensity outline, 4 – Pitch evolution, 5 – Unnormalized Formants $\widetilde{F}_1$ and $\widetilde{F}_2$, 6 – Spectrum.

The avatar has been developed with a static part (skull), and three dynamics components (tongue, jaw and lips). The normalized formant frequencies F_{kN} modify the horizontal and vertical positions of these components, based on the premises that $F1$ is correlated with the elevation of the tongue and $F2$ is correlated with tongue frontness (Watt et al., 2002).

As shown in the left hand side of Fig. 4, the tongue has two degrees of freedom whilst jaw has only one, whose coordinates are described by Eq. (1), where x_t, y_t and x_j, y_j are the Cartesian coordinates on screen position (in pixels) of the tongue and jaw respectively at rest stage, and α, β and γ are values (scale factors) obtained experimentally by the authors to set the units from formant frequencies in Hz, to units in pixels on the screen, so that they can be represented graphically on the screen. In this case $\alpha = 0.022$, $\beta = 0.063$ and $\gamma = 0.03$.

$$tongue\,(x_t + \alpha F_{2N}, y_t + \beta F_{1N}) \quad jaw\,(x_j, y_j + \gamma F_{1N}) \tag{1}$$

The lips model (right hand side of Fig. 4) has two independent degrees of freedom: one in the horizontal direction (point $p1$) located at the angle of the mouth, and other in the lower lip (affects $p5$ and $p6$). Points with notation px' means the same behavior of points px but on the other side of the mouth.

The behavior of the points $p1, \ldots, p6$ are governed by the expressions in Eqs. (2)–(4), where x_i, y_i with $i = 1, \ldots, 6$ are the Cartesian coordinates on screen position (in pixels) for each point, and Δx, Δy are the factors which moves the lips properly and are defined by Eq. (5). Eqs. (1)–(5) are proposed by the authors (experimentally) to model the behavior of these components (tongue, jaw and lips) of the avatar from formants frequencies estimated robustly, and to facilitate the avatar's representation on screen.

$$p1 = (x_1 + \Delta x, y_1) \tag{2}$$

$$p2 = (x_2, y_2), \quad p3 = (x_3, y_3), \quad p4 = (x_4, y_4) \tag{3}$$

$$p5 = (x_5, y_5 + \Delta y), \quad p6 = (x_6, y_6 + \Delta y) \tag{4}$$

$$\Delta x = k_1 \delta, \quad \Delta y = 0.85 \gamma F_{1N} \tag{5}$$

The value δ is the distance $\delta = \sqrt{F_{1N}^2 + F_{2N}^2}$ obtained from the normalized formant frequencies; the distance δ provides a relation of the distance between angles of the mouth. Rounded vowels like /o/ and /u/ have lower formant F1 values (lower δ) than open vowels do (higher δ). The scale factor k_1 fits the distance δ to the screen space in pixels, in this case $k_1 = 0.016$ and Δy is the vertical component of the lower lip; this value is proportional to the vertical component to the jaw.

The left hand side of Fig. 5 shows the complete joint model, and the right side the final user interface. The therapist selects the gender, height and vowel to train, and the system shows a pattern-vowel (thick line in Fig. 5-right) with the shape tongue for the vowel selected. This shape was built from combination of simple geometrical forms (three circles and lines) and magnetic resonance images MRI (Gurlekian et al., 2000) for each vowel. The goal is to try to match the own vowel-utterance to pattern-vowel.

Articula also provides useful additional information to the therapist about user's voice parameters such as intensity, pitch, formants (not normalized) and spectrum in real time, as shows Fig. 3 in items 3,4,5 and 6.

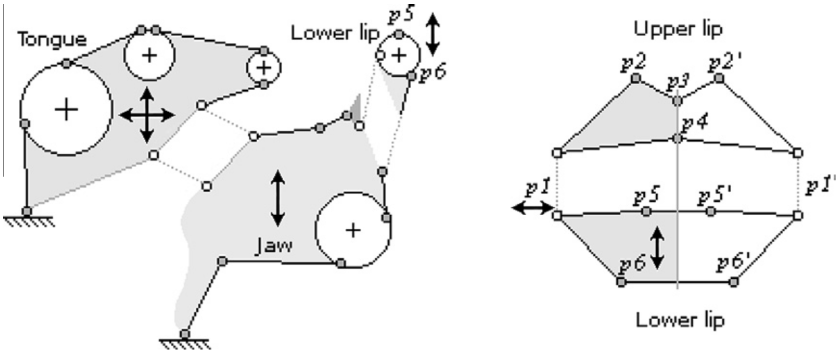


Fig. 4. Articula. Tongue and jaw models (left), lips model (right).

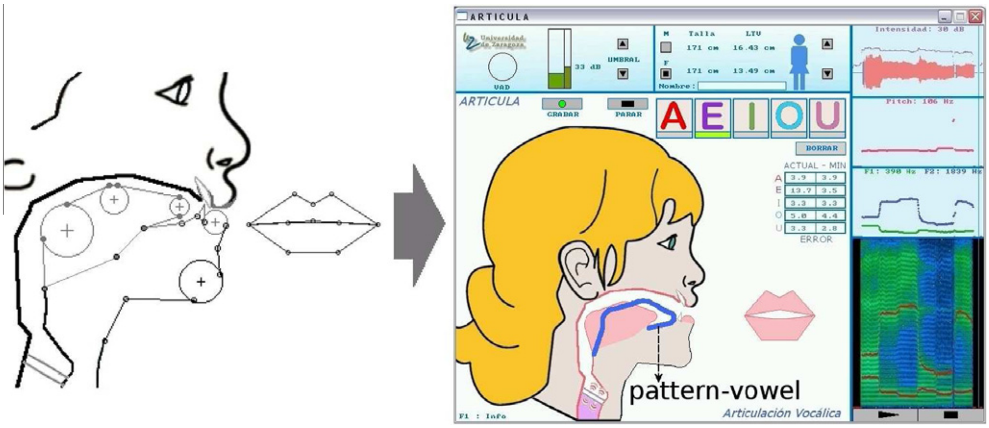


Fig. 5. Articula. Joint model (left) and final interface (right).

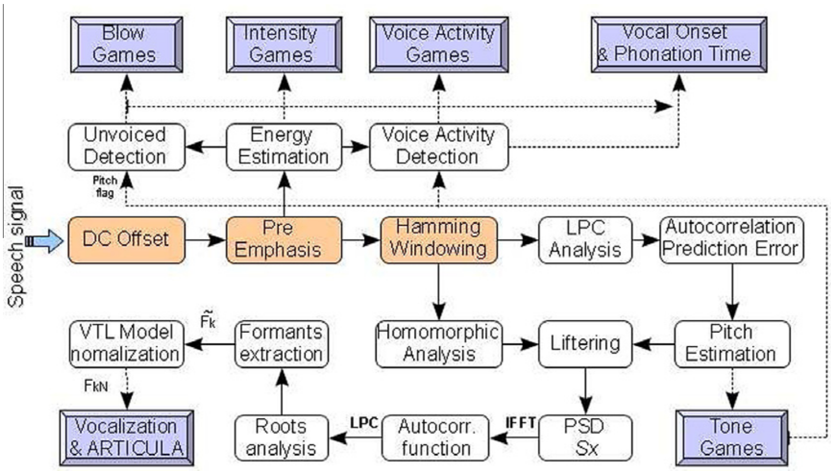


Fig. 6. PreLingua block diagram.

3. Speech technologies in PreLingua

As seen, *PreLingua* gathers activities to train patients (mainly children and severely disabled adults) with altered voice aiming to assist the work in speech therapy oriented to phonation. The corrective effect of this tool is based on the robustness of speech processing to know the values of different features in the user's voice, including intensity, tone, vocal onset, phonation time and, finally, formant frequencies.

A speech processing diagram block like the one shown in Fig. 6 is used for the correct estimation of all these features, and also show the corresponding game activity in *PreLingua*. In this diagram, after signal preprocessing (DC offset and pre-emphasis), the energy of each frame is calculated (required in *Intensity* activities) and a threshold is applied to determine whether voice from the user is presented on the frame or not (used for *Voice Activity* games). Signal is then windowed using a Hamming window and a LPC (Rabiner and Shafer, 1978) analysis is 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* activities) and sonority level as the ratio of voiced frames in a certain segment (used in the *Blow* activity). By combining voice activity detection and sonority estimation, it is possible to extract others as the *Vocal Onset* or the *Phonation Time*. *Vocalization* activity and *Articula* make use of formant estimation and normalization.

From all these features, the estimation of the pitch and the formants are those which require a bigger effort to provide these values robustly in the presence of different speakers, especially in those cases where a high value of pitch is present (children and women). A review on the traditional techniques to estimate these values (LPC and autocorrelation) will be made in following sections, as well as the study of more robust estimation methods for formant frequencies in Section 4.

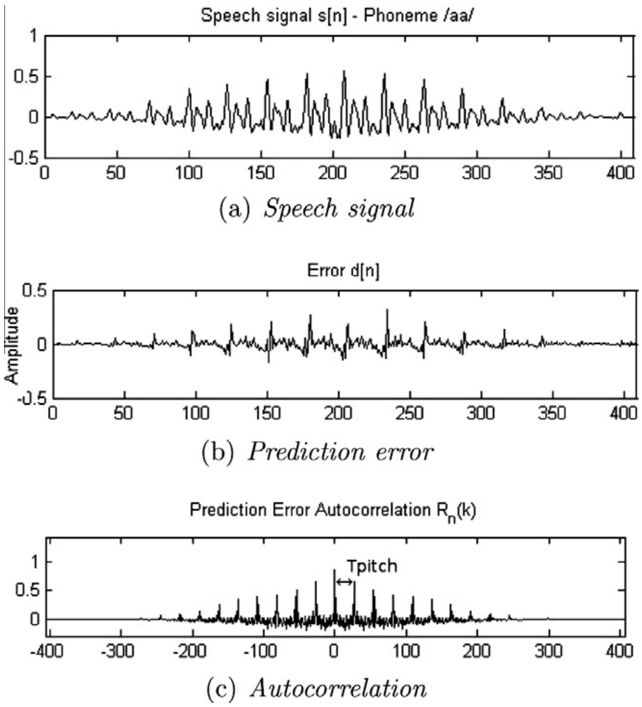


Fig. 7. LPC analysis and Tpitch estimation.

3.1. LPC analysis

LPC analysis is one of the most powerful and widely used techniques for speech analysis. The relevance of this method lies both in its ability to provide accurate estimates of the speech parameters and in its real-time computational speed. Based on the traditional model of speech production (Fant, 1960), the linear system is described by an all-pole system function relates the transform functions of the speech samples $s[n]$ and the excitation $e[n]$ via a gain, G and a set of filter coefficients a_i.

The basic problem of linear prediction analysis is to determinate the predictor coefficients a_i directly from the speech signal in order to obtain a useful estimate of the time-varying vocal tract system. The basic approach is to find a set of predictor coefficients that will minimize the mean-squared prediction error ϵ over a local segment of the speech waveform. The resulting parameters are then assumed to be the parameters of the system function $H(z)$ in the production model of the given segment of the speech waveform (Rabiner and Shafer, 2007). By calculating the autocorrelation of the input signal $s[n]$, it is possible to obtain the values of the parameters a_i via a Levinson–Durbin recursion.

3.2. Pitch estimation

A result of the LPC analysis is the generation of the error signal $d[n]$. To the extent that the actual speech signal is generated by a system that is well modeled by a time-varying linear predictor of order p, then $d[n]$ is equally a good approximation to the excitation source. Based on this reasoning, it is expected that the prediction error will give information (for voiced speech) about the periodicity of the excitation source. The pitch period can be estimated by performing an autocorrelation analysis on $d[n]$ and detecting the largest peak in the appropriate range. Rabiner and Shafer (1978). Fig. 7 shows a speech segment of the phoneme /a/ after Hamming windowed in Fig. 7(a), the error $d[n]$ from LPC analysis in Fig. 7(b) and the autocorrelation of the prediction error $R_n[k]$ in Fig. 7(c), where the lag value of the largest peak detected corresponds to the pitch period *Tpitch*. The pitch frequency is obtained from $F_p = 1/Tpitch$ and its evolution in time is used in *Tone* games.

3.3. Formant estimation

From the extraction of LPC parameters, the system can be described as the output of a FIR linear ($A(z)$) system in terms of its zeros. According to this, the zeros of $A(z)$ are the poles of $H(z)$ and, therefore, it can be expected that roughly $\frac{Fs}{1000}$ of the roots will be close in frequency (angle in the z-plane) to the formant frequencies (with Fs the sam-

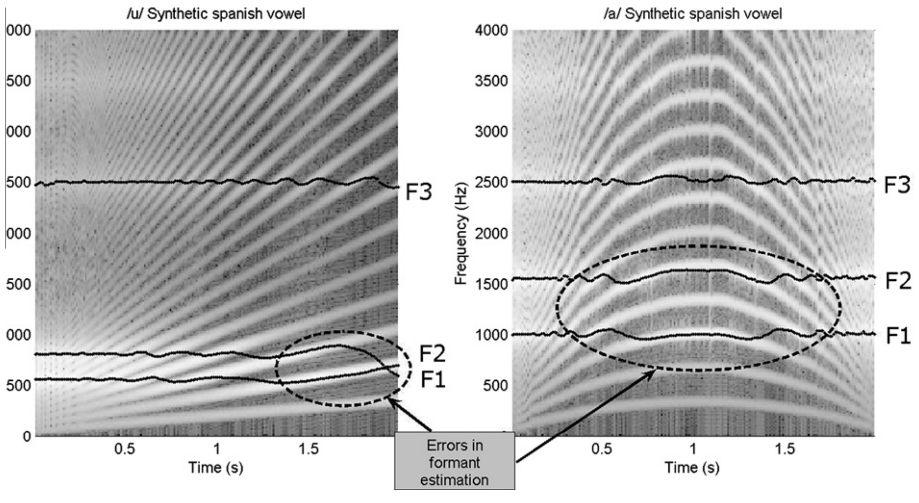


Fig. 8. High-pitch influence in formant estimation for synthetic /u/ and /a/ Spanish vowels.

pling frequency in Hz). That is, the roots (complex conjugate pairs) that are close to the unit circle, are the poles of $H(z)$ that model the formant resonances (Rabiner and Shafer, 2007).

By knowing the formant frequencies $F1$ and $F2$ it is possible to identify the vowels in order to build vocalic articulation activities in *PreLingua*; but the model of LPC analysis suffers from difficulties in estimating vocal tract characteristics of high-pitched speakers, which limits its application in tools for voice therapy. Hence, it was necessary to apply more complex techniques in speech processing to try to solve this problem, as it will be seen in the next Section.

4. Formant estimation and normalization

The formant measurement is technically difficult. The situation is less severe in male adult cases in which the fundamental frequencies (F0) are low (Traunmuller and Eriksson, 1997). In female and children cases F0 increases, hence F0 and its harmonics could match or be very near to the formant values, affecting the estimation (Rodríguez and Lleida, 2009).

4.1. Robust formant estimation in high pitch voices

The conventional autocorrelation method of linear prediction LPC, works well in signals with long pitch period (low-pitched). As the pitch period of high-pitched speech is small, the periodic replicas cause aliasing in the autocorrelation sequence. In other words, the accuracy of the LPC method decreases as the fundamental frequency F0 of speech increases. Fig. 8 shows the formant estimation for synthetic Spanish vowels (/u/ and /a/) using LPC method with order $p = 8$, over 25 ms long speech frame. The filter coefficients for the all-pole vocal tract model are obtained through Durbin's recursion using the autocorrelation method, after Hamming-windowed the pre-emphasized ($\alpha = 0.97$) speech frame. When F0 increases the formant estimation tends to the pitch harmonics (dashed ellipse), situation which hides the real value of the formant. In that case, it is required to separate these effects in order to obtain formants not contaminated by $F0$.

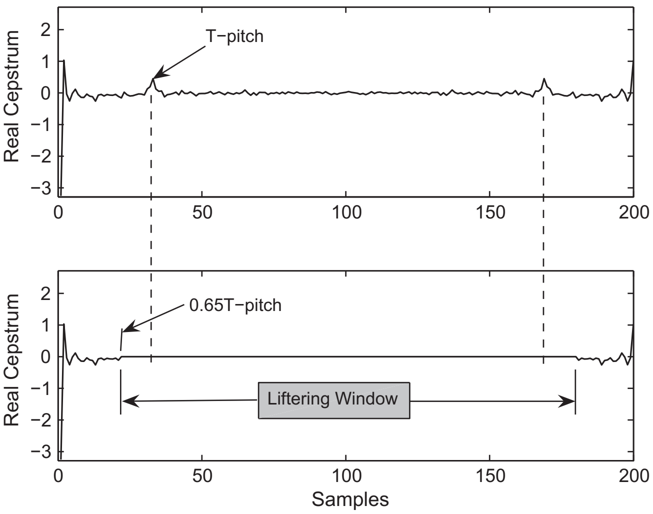


Fig. 9. Effect of liftering in the real cepstrum domain.

An alternative to reduce the aliasing in the autocorrelation sequence is to use the homomorphic analysis as proposed in (Shahidur and Shimamura, 2005). The main idea within the homomorphic analysis is the deconvolution of a segment of speech $x[n]$ into a component representing the vocal tract impulse response $e[n]$, and a component representing the excitation source $h[n]$.

The way in which such separation is achieved is through linear filtering of the cepstrum, defined as the inverse Fourier transform of the log spectrum of the signal. As the cepstrum in the complex domain is not suitable to be used because of its high sensitivity to phase (Rabiner and Shafer, 1978), the real-domain cepstrum $c[n]$ defined by Eq. (6) is used, where $X(k)$ is the N-point Fourier transform of the speech signal $x[n]$.

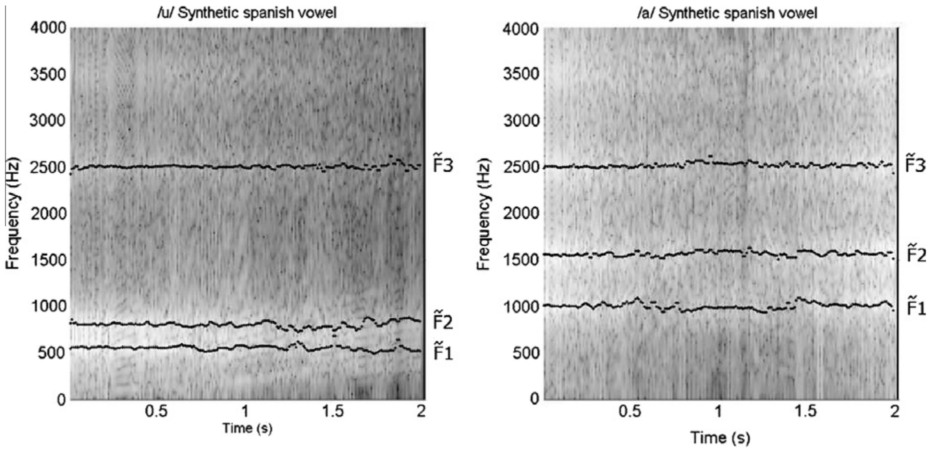


Fig. 10. Synthetic vowels /u/ and /a/ after liftering process.

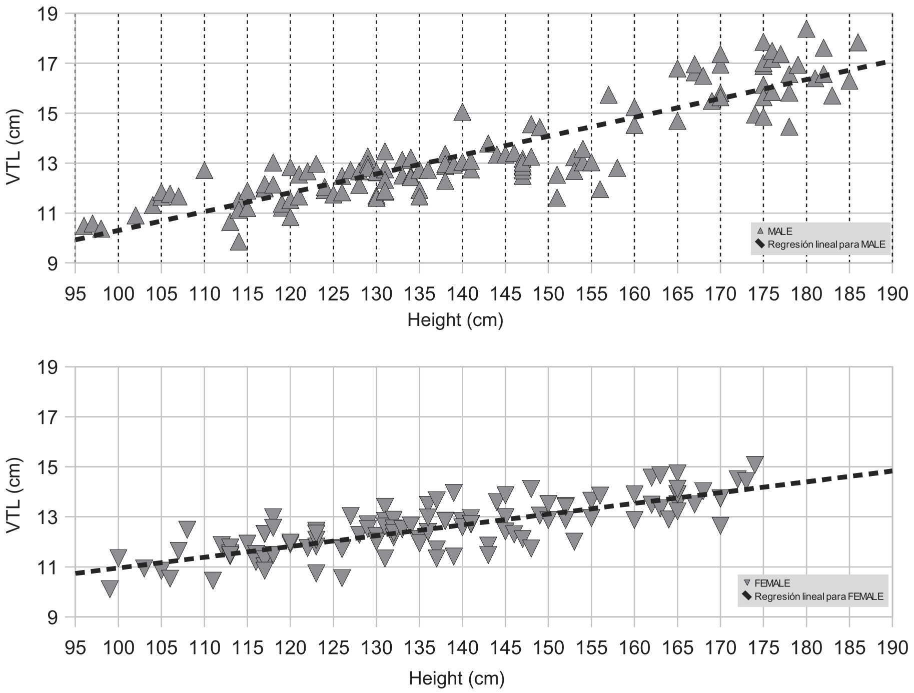


Fig. 11. VTL Estimated for 235 children 110 (female) and 125 male (up).

$$c[n] = \frac{1}{N}\sum_{k=0}^{N-1} ln|X(k)|e^{j\frac{2\pi}{N}kn}, \quad 0 \leqslant n \leqslant N-1 \tag{6}$$

The values of $c[n]$ around the origin correspond primarily to the vocal tract impulse information, while the farthest values are affected mostly by the excitation. Knowing previously the value of the pitch period T_{pitch} from the LPC analysis using the autocorrelation method it is possible to filter the cepstrum signal (liftering) and use the liftered signal to find the formant frequencies. A liftering window with length of $0.5T_{pitch}$ has been proposed in (Verhelst and Steenhaut, 1986) or $0.6–0.7T_{pitch}$ in (Shahidur and Shimamura, 2005). In this work, the liftering window $w[n]$ in Eq. (7) is $0.65T_{pitch}$ and the effect of applying $w[n]$ in the real cepstrum domain can be observed in Fig. 9.

$$w[n] = [0.65T_{pitch}, N-1-0.65T_{pitch}] \tag{7}$$

After the liftering process, the formant frequencies $\widetilde{F}_k$ without pitch influence are obtained through similar LPC method described above, with order $p = 8$ and 25 ms speech frame. Fig. 10 shows the same vowels of Fig. 8 after the liftering process where the effect of the pitch and its harmonics are removed.

4.2. Vocal tract length estimation

This section will describe a robust method to estimate the vocal tract length from formant frequencies. The length can be estimated from the formant information of the user, and it is based on the model of the vocal tract as a uniform lossless tube. Modeling the vocal tract as a uniform lossless acoustic tube, its resonants frequencies given by Eq. (8) are uniformly spaced, where $v = 35300$ cm/s is the speed of sound at 35°C, and l is the length of the uniform tube in cm.

$$F_k = \frac{v}{4l}(2k-1), \quad k = 1, 2, 3, \ldots \tag{8}$$

The estimation of the length was proposed in (Necioglu et al., 2000), and it can be reduced to fitting the set of resonance frequencies of a uniform tube, which are determined solely by its length l. Therefore, the problem can be approximated to minimizing Eq. (9), where $D(\widetilde{F}_k, (2k-1)F1)$ is a function that express the difference between the measured formant $(\widetilde{F}_k)$ and the resonance of the uniform tube.

$$\varepsilon = \sum_k D(\widetilde{F}_k, (2k-1)F_1) = \sum_k D\Big(\widetilde{F}_k, (2k-1)\frac{v}{4l}\Big) \tag{9}$$

From Necioglu et al. (2000), the error measure given in Eq. (9) can be turned in Eq. (10) using the distance function between the measured formant $\widetilde{F}_k$ $(k = 1, \ldots, M)$ and the odd resonances of a uniform tube, $(2k-1)F_1$.

$$\varepsilon = \sum_k \frac{\left(\frac{\widetilde{F}_k}{2k-1} - F_1\right)^2}{F_1} \tag{10}$$

Finally, minimizing Eq. (10) into Eq. (11) in order to obtain the estimated resonance frequency of the uniform tube (F_1), the vocal tract length VTL can be obtained with the expression in Eq. (12).

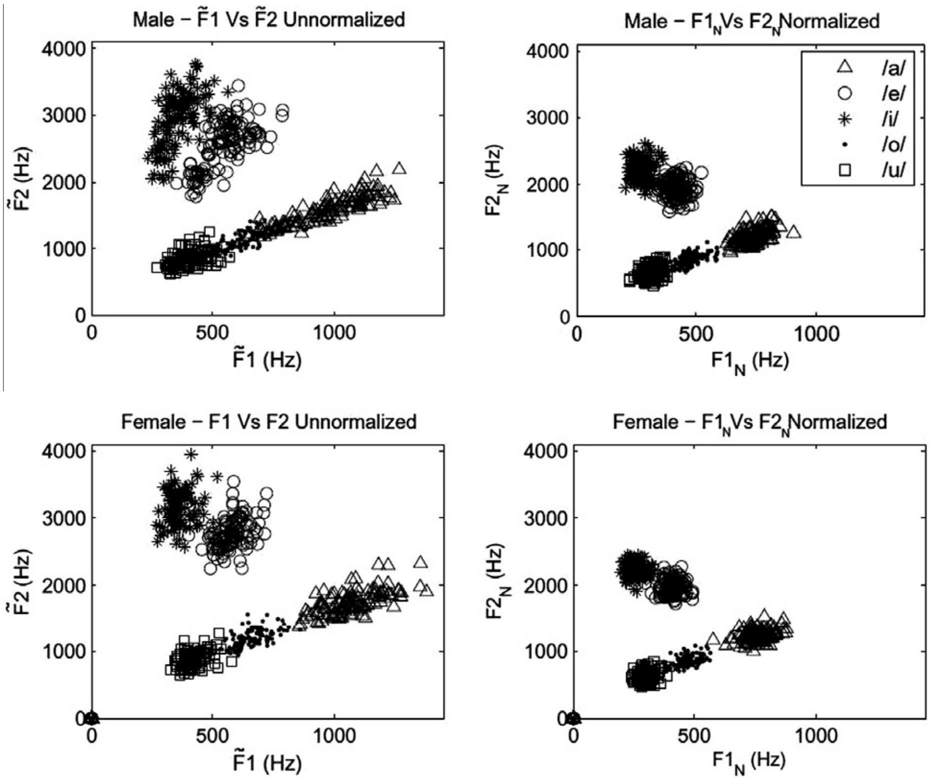


Fig. 12. Comparison between $\widetilde{F}_k$ and F_{kN}, male (up) female (down) for Spanish vowels.

$$F_1 = \left(\frac{1}{M}\sum_k\left(\frac{\widetilde{F}_k}{2k-1}\right)^2\right)^{1/2} \quad (11)$$

$$VTL = \frac{v}{4F_1} \quad (12)$$

This method was applied in a previous study (Rodríguez and Lleida, 2009) where, through speech recordings from 235 healthy children of different ages, it was possible to find out a correlation between the vocal tract length and the height in children from 3 to 17 years old, as shown in Fig. 11.

4.3. Formant normalization

With the formant frequencies $\widetilde{F}_k$ obtained in Section 4.1, we can normalize these estimations through the vocal tract length estimated in Section 4.2. The formant normalization used in this study has been proposed by Wakita (1977). That work is based on the hypothesis that the vocal tract configuration of the speakers are similar to each other and differ only in length. Based upon the hypothesis for normalization, it is necessary to compute the resonance frequencies of an acoustic tube when the length of the tube l is varied to a reference length l_R without altering its shape. Hence, the normalized formants F_{kN} are computed by multiplying the unnormalized formants $\widetilde{F}_k$ by the length factor, l/l_R, with l_R fixed at 17.5 cm. As shown in Eq. (13), l corresponds to the vocal tract length VTL obtained from Section 4.2.

$$F_{kN} = \frac{l}{l_R}\widetilde{F}_k \quad (13)$$

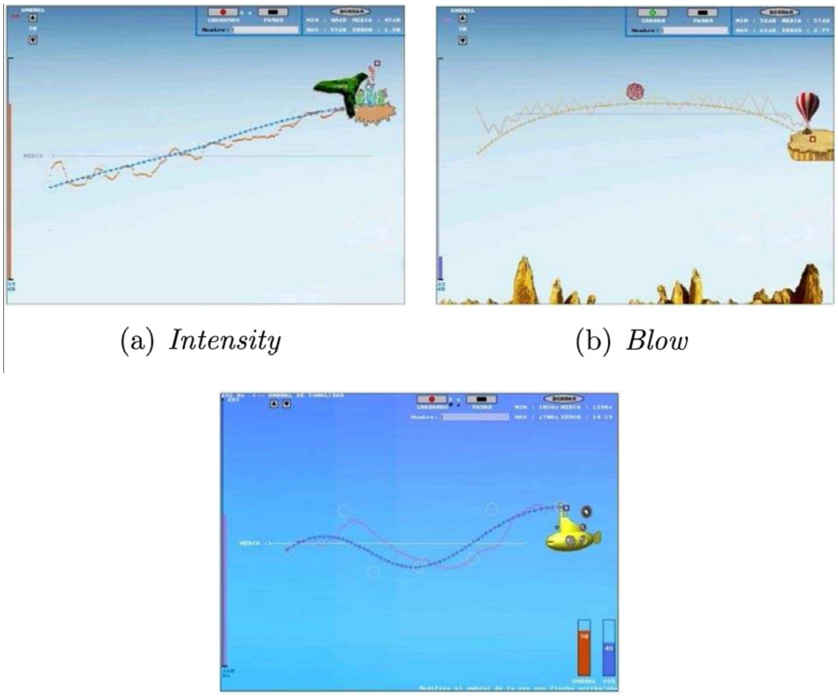
A graphic comparison between unnormalized formants $\widetilde{F}_k$ and normalized formants F_{kN} can be appreciated in Fig. 12. This figure shows the five Spanish vowels from 235 healthy children (125 male (up), 110 female (down)) before normalization with high dispersion (left) and, how the dispersion due to inter-speaker variability is reduced after the normalization process (right).

5. Experimental study with PreLingua

In order to provide an objective assessment on the phonatory faculties of a given patient, *PreLingua* offers a section to evaluate three different features like intensity, blow and tone in individual sessions. In the *Articula* activity the system also evaluates each vowel for each session if the therapist decides so. This evaluation section is located at the right hand side of the *PreLingua* main window (Fig. 1). The evaluation consists in reports which show statistical data about the session in a readable format and an image containing a screen-shot of the activity. The therapist can use the reports and images to complement the clinical information.

Each evaluation activity allows the therapist to define the therapy targets (pattern) depending on the level of

Please cite this article in press as: Rodríguez, W.R. et al. A prelingual tool for the education of altered voices, Speech Comm. (2011), doi:10.1016/j.specom.2011.05.006



(a) *Intensity*

(b) *Blow*

(c) *Tone*

Fig. 13. Evaluation activities. Pattern in dotted line and actual production by the patient in continuous line.

performance of the user. The therapist can explain to the patient the way to carry on the activity, for instance with his/her own voice production. When the activity is finished, the system generates the report which contains information about the Minimum Value (MINV) across the session, the Maximum Value (MAXV) across the session, the dynamic range (MAXV–MINV), the mean, the Mean Square Error (MSE) between the pattern and the production of the user, the duration of the session in seconds and the date and time in which the session was carried out. At the same time the system saves graphical information as an image. Speech therapists work in sessions whose duration depend on the characteristics and abilities of the patients and varies largely from one patient to another, the duration of a session may range from 20 to 60 min according to patient's skills.

Fig. 13 shows the 3 evaluation activities proposed in *PreLingua*. Fig. 13(a) presents a rising pattern for *Intensity* and the performance of a patient, Fig. 13(b) shows a tableland pattern for *Blow* and the performance of the patient; and, finally, Fig. 13(c) shows a rising-falling pattern for *Tone* and the performance of a given patient. As mentioned, *Articula* can also store the evaluation of the vowel production for each session, consisting in measuring the MSE between the vowel pattern provided by the system and the position of the avatar's tongue as moved by the vowel produced by the user as shown in the right hand side of Fig. 5.

This feature permitted to make an evaluation of the capabilities of *PreLingua* as a tool to improve and assess the phonatory and articulatory abilities of patients with voice alterations. *PreLingua* was tested for three months in the Center of Special Education "CEDESNID' '(CEDESNID) in Bogotá (Colombia) and the Public School for Special Education (CPEE) "Alborada" (La Alborada) in Zaragoza (Spain). The test was applied in users with different disabilities (cognitive and/or motor) that affected their voice skills in many different ways.

Hence, the study which was applied in this work focused on a group of patients from the two educative institutions and aimed to compare their voice skills at the beginning of the study and their voice skills after 12 weeks of therapy with *PreLingua*. The study started with 39 subjects, but only 27 finished it due to the continuous absences of the patients in the sessions due to several causes. The 27 subjects were found as potential users of the system, which limited the possibility to establish a control group which followed the traditional speech therapy techniques for comparison with the novel tool. In the end, it was decided that all the users took advantage of the capabilities of *PreLingua*. Table A.1 shows the characteristics of the group of subjects: Gender, age, location and the overall diagnosis provided by their therapists for each user. All of the 27 users participated in the activities of intensity, blow and tone; while 24 of them participated in the activities with *Articula*. The user selection was made by the therapists based on each user capability of understanding the proposed therapy.

At the initial stage of the study, each user was evaluated by the therapist through a speech therapy evaluation form in order to assess their voice skills; afterwards, the therapists used *PreLingua* for 12 weeks with each user; and, finally, the therapists repeated the same speech therapy evaluation with each user to assess changes in the voice skills of the subjects. The therapists were free to decide how to use *PreLingua* with each subject, focusing on those skills which might be more helpful or necessary for the patient.

The objective assessment of the subjects was made with data from the first 3 weeks of therapy with *PreLingua* and from the last 3 weeks of therapy with *PreLingua*. Subjects performed the evaluation activities described previously and the results in intensity, blow, tone and *Articula* were stored for a posterior analysis. Only the MSE between the proposed pattern and the user's performance was considered; and the mean value and standard deviation of this value was calculated for the initial and final sessions separately. The final aim was to detect differences between the results of the activities at the beginning and the end of the therapy to detect possible significant changes in the subjects abilities.

6. Results

The results described here are divided in objective measurements from the statistical analysis of the results stored by *PreLingua* and subjective mesurements from the pre and post speech therapy evaluation forms for each user.

6.1. Objective assessment with PreLingua

Table A.2 shows the results on how the different subjects of the study achieved improvements in their voice production for the different activities (intensity, blow, tone and *Articula*) according to the values of MSE stored internally by the application. The improvement (*Imp.*, Yes or No) was consider as existing (Yes) in cases with reduction in the average MSE between the mean of first three sessions and the mean of last three sessions with significance above 50% ($p < .5$), and was consider as negative (No) in cases without reduction in the average MSE. As mentioned previously, the MSE measured the distance between the pattern of intensity, blow, tone or pattern-vowel proposed to the user and his/her actual production of voice.

Based on the data obtained by the system, a t-test of statistical significance was applied for each user in order to establish if the improvement on the subject's voice skills was truly significant or it just happened by chance. The values of the MSE at the beginning and the end of the test (X_1 and X_2) were characterized with their means and standard deviations ($\overline{X_1}, \overline{X_2}, \sigma_1$ and σ_2) and the number of samples in each case (n_1 and n_2) which was 3, as 3 sessions were analyzed at the initial and final stages of therapy. As the variances of the data to compare were unequal, an adaptation of Student's t-test called Welch's test was used.

For the significance test, the distribution of the statistics was approximated as being an unpaired 2-sided Student's t distribution with the degrees of freedom obtained by the Welch–Satterthwaite equation.

There is no specific or determined level of significance which can serve to make the significance of the results provided, as users have different levels of disabilities and the length of the study could not be as long as desired. The significance results provided in Table A.2 were very variable for each activity and patient. By considering a significance level ⩾99% ($p < .01$) the study showed improvements in 4 subjects (14.8% of the total) in intensity, 5 subjects (18.5%) improved in blow, only 1 subject (3.7%) in tone and 2 subjects (8.3%) improved in at least one vowel. A more relaxed threshold, but still determining high significance (⩾95% ($p < .05$)) provided improvements in 8 subjects (29.6% of the total) in intensity, in 7 patients (25.9%) in blow, in 6 subjects (22.2%) in tone and 5 users (20.8%) in at least one vowel. Finally, a level of significance ⩾80% ($p < .2$) marked improvements in 15 subjects (55.6%) in intensity, in another 15 (55.6%) in blow, in 8 users (29.6%) in tone and in 8 users (33.3%) in at least one vowel. It is to be remembered again that 27 subjects participated in intensity, blow and tone evaluation and 24 in the vocalic evaluation.

In general, intensity and blow were the activities where more subjects achieved significant improvements at all the significance levels. These results were especially encouraging for blowing activities, as they require a higher level of concentration compared to the intensity activities which can be considered to be easier. Less subjects achieved significant improvements in tone, possibly influenced by the short duration of the therapy study, as subjects require a high level of awareness to improve in this feature and more time would be required to achieve a better control of the vocal folds by the patient.

Regarding the vocalic articulation activity, it is well known that the articulation process is affected by geometrical features of the patient's vocal cavity. Some of the subjects presented malformations in the hard and soft palates, crooked teeth and/or hypotonia or hypertonia. So the achieved results were not as relevant as for the other activities and this voice skill would also require extended therapy sessions.

Comparing across different vowels, a small number of subjects showed improvements in the articulation of vowels /a/ and /o/, whose first formant $F1$ is higher than for other vowels (only 2 significant cases), highlighting the difficulty opening the mouth by the users. On the contrary, the improvements in articulation for vowels /e/, /i/ and /u/ achieved better levels of significance, since the effort required in opening the mouth is less.

6.2. Subjective assessment by the therapists

Regarding the observations provided by the therapists in their pre and post therapy assessment, the Table A.3 summarizes the 27 cases of study with the speech therapy evaluations before and after applying *PreLingua*. The criteria of the therapists are to be able to work in pre-linguistic communication and voice skills as: Intensity, Blow Duration, Tone, Tongue Praxis, Rhythm, and finally, an additional column with topics highlighted by the therapist as additional skills observed. The speech therapy evaluation used was created by speech therapists under the patronage of the "Junta de Andalucía", Regional Government of Andalucía (Spain), and was recommended for the use in the study by one of the collaborating therapists in the CPEE "Alborada"; it included many aspects in language acquisition like auditory, visual and attention skills; voice features like intonation, blowing, rhythm and intensity; and anatomical aspects and praxis. All of them were evaluated in subjective scales according to each topic for evaluation; for instance, rhythm was evaluated as normal, tachylalia, gasped or bradylalia; or tone was evaluated as normal, monotonous or robotic. In Table A.3, a summary column is included for each feature marking whether a positive change was seen comparing the pre and post therapy assessments. The caption of Table A.3 shows also how all these features were evaluated.

Blow duration was the skill where a larger number of subjects had an improvement according to the therapists, followed by intensity and rhythm. In the intensity feature, 12 subjects had a positive change after therapy, for instance changing from strained voice to normal voice (although with some difficulties) or increasing skill in asthenic voices. 18 subjects improved in blow duration, corresponding to cases like subjects 1, 2, 3, 4 and 7 among others, who changed from normal blow to normal with increased skill ($N \Rightarrow N, I - S$), or subject 15 who changed from altered blow to normal with difficulties ($AL \Rightarrow N, WD$). Regarding tone, only 7 subjects suffered a positive evolution of their skills, most of them getting closer to a normal intonation compared to their previous monotonous intonation. Tongue praxis was the skills where *Articula* acted during therapy, with 8 cases improving their performance for the therapists; in all cases increasing their skills in the tongue movements. The rhythm aspect was treated by activities in *PreLingua* like Vocal Onset and Phonation Time, 12 subjects showed an improvement in this skill, most of them achieving a normal rhythm.

Finally, the last column in Table A.3 captures all those additional skills observed by the therapist to have been acquired by the patients. Out of the 27 cases of study, 21 of them achieved skills which were not to be considered as an outcome of the study. These skills included increases in attention time, higher ability to follow instructions, better control of blow direction and socialization skills.

At the end of the experience, the therapists were asked to evaluate the work with *PreLingua*. They considered it to be very easy to use and very attractive for all the patients (disabled children and adults). They observed improvements during the 12 weeks of work with *PreLingua*, highlighting the ability of sustained blowing instead of disrupted blowing and better continuous patterns (tableland and rising–falling) in voiced utterances. As additional features of applying *PreLingua* in this population, the therapists mentioned the arising possibilities of applying the tool in different areas related to special education like hearing impaired users, muteness, cases of stroke, autism and apraxia.

Some other observations mentioned cognitive issues, as some subjects showed improvements in paying attention, better levels of concentration and memorization and a high motivation of the user. Regarding senso-perceptual issues, some subjects showed better spatial location (on the screen) and coordination, as well as improvements in visual and auditory perception. In communication skills some subjects showed an increase in voiced utterances and the recognition of the different characters inside *PreLingua*. For instance in a case reported outside of this study, a child with deep cognitive delay without oral communication achieved voiced utterances after continuous therapy with levels 1 and 2 of the activities in *PreLingua* (voice detection and intensity). An issue often mentioned by the therapists was related to socialization skills, as they mentioned positive attitudes like team playing, taking turns to play, helping each other, healthy competition and auto demanding in some cases.

Concerning *Articula*, the therapists also evaluated this application as friendly and attractive for all users, highlighting the appropriate and apprehensible interface to train vocalic articulation in real time, ideal for motivating children in speech therapy.

7. Discussion

The main points of discussion which arose after the examination of the results in the previous Section were, on one hand, to determine if *PreLingua* could be considered as a successful tool for improving the voice skills of patients with alterations in their voice and, on the other hand, to see if *PreLingua* and its activities could also serve to assess these skills in different users along time. As it was mentioned previously, the work with these subjects was complicated as they presented many different alterations in their vocal tract which made extremely difficult their speech and any therapy with them.

The observations from the therapists and the improvements shown by a number of the subjects who participated in the study using *PreLingua* during 12 weeks indicated that the tool had certainly improved the voice skills of many of them. Elements like blow, intensity and rhythm, which were improved after the therapy by 66.6%, 44.4% and 44.4% of the subjects, respectively, according to the therapists' assessment, were largely influence by the continuous work with *PreLingua*. Those skills like correct intonation and vocalic articulation which require a better control of muscles and physiological parts of the vocal tract were not affected in as many as the users as the previously mentioned skills. Only 25.9% and 33.3% of the subjects were noticed by the therapists to improve in these skills

Table A.1
Descriptions of the patients used in the experimental study.

Case	Gender	Age	Location	Diagnosis
1	Male	14	Colombia	Moderate mental delay, cerebral palsy
2	Male	18	Colombia	Mild mental delay, cerebral palsy
3	Male	13	Colombia	Communication disorder
4	Male	17	Colombia	Moderate mental delay, cerebral palsy
5	Male	22	Colombia	Moderate mental delay, comm. disorder
6	Male	18	Colombia	Moderate mental delay, cerebral palsy
7	Male	20	Colombia	Moderate mental delay, convulsive synd.
8	Male	34	Colombia	Moderate mental delay, Down synd.
9	Male	18	Colombia	Moderate mental delay, Convulsive Synd.
10	Male	24	Colombia	Communication disorder
11	Male	23	Colombia	Severe mental delay
12	Female	18	Colombia	Moderate mental delay
13	Male	18	Colombia	Severe mental delay
14	Male	17	Colombia	Moderate mental delay
15	Female	34	Colombia	Moderate mental delay
16	Male	13	Colombia	Severe mental delay
17	Male	14	Colombia	Moderate mental delay
18	Male	17	Colombia	Moderate mental delay
19	Male	21	Colombia	Moderate mental delay
20	Female	21	Colombia	Moderate mental delay
21	Male	12	Colombia	Moderate mental delay
22	Male	11	Spain	Moderate mental delay, hypertonia
23	Female	16	Spain	Moderate mental delay, hypotonia
24	Male	15	Spain	Moderate mental delay, Quadriplegia
25	Female	14	Spain	Moderate mental delay
26	Female	14	Spain	Moderate mental delay
27	Female	16	Spain	Moderate mental delay, Down Synd.

Table A.2
Objective measurements for Intensity, Blow, Tone and Articula activities for each user. Yes: improvement or reduction in the MSE between the first three sessions and the last three sessions, No: not improvement or not reduction in the MSE.

Case No.	Intensity		Blow		Tone		Articula	
	Imp.	$p < XX$	Imp.	$p < XX$	Imp.	$p < XX$	Imp. Vowel	$p < XX$
1	Yes	0.23	Yes	0.15	No	–	/u/	0.41
2	Yes	0.04	No	–	Yes	0.2	No	–
3	Yes	0.7	No	–	No	–	/e/	0.28
4	Yes	0.33	No	–	No	-	N/A	N/A
5	No	–	No	-	Yes	0.21	N/A	N/A
6	No	–	No	-	Yes	0.21	/i/ /o/	0.19 0.37
7	Yes	0.24	No	–	No	-	N/A	N/A
8	Yes	0.86	No	–	No	–	No	–
9	Yes	0.13	Yes	0.19	No	–	No	–
10	Yes	0.2	Yes	0.16	Yes	0.2	No	–
11	Yes	0.87	Yes	0.25	Yes	0.15	No	–
12	Yes	0.17	No	–	No	–	No	–
13	Yes	0.14	Yes	0.31	Yes	0.36	No	–
14	Yes	0.29	No	–	Yes	0.34	No	–
15	Yes	0.01	Yes	0.01	No	–	/o/	0.1
16	Yes	0.01	Yes	0.05	Yes	0.08	/i/ /u/	0.11 0.07
17	Yes	0.59	Yes	0.06	Yes	0.15	/e/	0.39
18	Yes	0.96	Yes	0.08	Yes	0.36	/a/	0.22
19	Yes	0.7	Yes	0.13	No	–	/e/	0.26
20	Yes	0.29	Yes	0.05	No	–	No	–
21	Yes	0.01	Yes	0.44	Yes	0.3	No	–
22	No	–	Yes	0.99	No	–	/u/	0.31
23	Yes	0.21	Yes	0.13	No	–	/a/ /i/	0.32 0.88
24	No	–	No	–	No	–	/a/ /e/ /u/	0.41 0.23 0.28
25	Yes	0.31	Yes	0.14	No	–	/o/ /u/	0.05 0.32
26	No	–	Yes	0.13	No	–	/i/	0.32
27	No	–	Yes	0.11	No	-	No	–

Please cite this article in press as: Rodríguez, W.R. et al. A prelingual tool for the education of altered voices, Speech Comm. (2011), doi:10.1016/j.specom.2011.05.006

Table A.3
Subjective results. Speech therapy evaluations before and after the study. A: Asthenic, AL: Altered, BD: Blow Direction, BR: Bradylalia, CN: Can Not, D: Decrease, FI: Follow Instructions, G: Gasped, I: Increase, IA: Increase Attention Time, M: Monotonous, N: Normal, R: Robotic, S: Skill, SS: Socialization Skill, ST: Strained, TL: Tachylalia, WD: With Difficulty, WE: With Effort, Add: Additional skills observed.

Case No.	Intensity			Blow Duration			Tone			Tongue Praxis			Rhythm			Other
	Before	After	Imp.	Before	After	Imp.	Before	After	Imp.	Before	After	Imp.	Before	After	Imp.	
1	N	N	No	N	N, I-S	Yes	N	N	No	CN	I-S	Yes	G, WD	I-S	Yes	IA, FI, SS
2	N	N	No	N	N, I-S	Yes	N	N	No	CN	I-S	Yes	N, WD	I-S	Yes	SS
3	N	N	No	N	N, I-S	Yes	R	N	Yes	N	N	No	N, G	I-S	Yes	IS, EI, SS
4	A	N	Yes	N	N, I-S	Yes	N	N	No	–	–	–	G, WD	G, WD	Yes	IA
5	A	A	No	N	N	No	M	M-N	Yes	–	–	–	BR	N	Yes	–
6	A	N	Yes	AL	AL, N-WE	Yes	M	N	No	N	N	No	N, WD	N	Yes	IA, FI, SS
7	N	N	No	N	N, I-S	Yes	N	N	No	–	–	–	G	N	Yes	IA
8	A	N	Yes	AL	AL	No	N	N	No	N	N	No	N	N	No	–
9	A	A	No	N	N, I-S	Yes	M	M	No	WD	WD	No	G, WD	N	Yes	FI, SS
10	N	N	No	N	N, I-S	Yes	N	N	No	N	N	No	N, WD	N	Yes	–
11	N	N	No	AL	AL	No	M	M	No	WD	WD	No	G	G	No	IA, SS
12	ST	N-ST	Yes	AL	AL	No	M	M	No	N	N	No	N	N	No	SS
13	AL	AL	No	AL	AL	No	M	M	No	N	N	No	BR	BR	No	FI, IA
14	ST	N-WE	Yes	N	N	No	M	N-WE	Yes	N	N	No	TL	TL	No	–
15	ST	N-WE	Yes	AL	N-WD	Yes	N	N	No	N	N	No	TL	TL	No	SS
16	ST	ST	No	AL	AL	No	M	M	No	N	N, I-S	Yes	TL	TL	No	SS
17	N	N	No	AL	AL, I-S	Yes	N	N, I-S	Yes	N	N	No	N	N	No	BD
18	AL	AL	No	AL	AL	No	M	M	No	N	N	No	TL	TL	No	FI
19	ST	ST	No	N	N	No	N	N	No	N	N	No	TL	TL, N-WE	Yes	BD
20	AL	AL, N-WD	Yes	AL	AL, N-WE	Yes	M	M	No	N	N	No	TL	TL	No	–
21	A	N-WE	Yes	AL	AL, N-WE	Yes	N	N	No	N	N	No	N	N	No	BD, SS
22	ST	N-WE	Yes	N	N, I-S	Yes	N	N	No	WD	N, WE	Yes	TL	TL	No	IA, SS
23	A	N-WE	Yes	N	N, I-S	Yes	M, R	M, N-WE	Yes	WD	WD, I-S	Yes	G	G	No	FI, IA
24	N	N	No	AL	AL, I-S	Yes	N	N	No	WD	WD, I-S	Yes	N	N	No	IA, SS
25	A	A, I-S	Yes	AL	AL, I-S	Yes	M	M, N-WE	Yes	N	N, I-S	Yes	N, G	N	Yes	SS
26	A	A, I-S	Yes	AL	AL, I-S	Yes	M	M, N-WE	Yes	N	N, I-S	Yes	G	G, N-WE	Yes	–
27	N	N	No	N	N, I-S	Yes	N	N	No	N	N	No	N	N	No	FI, BD

respectively. As these speech features were harder for the subjects to train, a further study which reflected variations through longer periods of time (half a year or a year) would be required to know if *PreLingua* has similar properties to improve these skills in individuals with altered voices as it improved intensity, blow and rhythm.

Therapists observed that *PreLingua* had a high motivational power to attract the attention of the subjects involved in the study (children and adults with mid to severe cognitive disabilities). The special interface of *PreLingua*, designed with a friendly environment, made it very successful with the disabled users of the study despite their ages. It should be evaluated if an unimpaired adult who might have lost voice due to trauma or disease would be so appealed by this interface. Although the performance of the techniques within *PreLingua* would be a priori as useful in unimpaired adults as with the impaired peers, a certain redesign of the user interface would be required to make it more suitable for adult population.

A very interesting part of the study came when comparing the subjects that improved skills according to the therapists and the improvements that these cases had experienced in the evaluation activities in *PreLingua*. 9 out of the 12 subjects with improvements in intensity

Please cite this article in press as: Rodríguez, W.R. et al. A prelingual tool for the education of altered voices, Speech Comm. (2011), doi:10.1016/j.specom.2011.05.006

according to the therapists also performed better in the final evaluation activities with different levels of significance. Similarly 12 out of 18 subjects that improved blow also did better in the application by the end of the therapy. 8 subjects were positively reviewed to have better praxis of their tongues after the therapy and 7 of them performed better with *Articula* in, at least, one of the 5 Spanish vowels. Tone was the only feature studied with *PreLingua* were the positive cases assessed by the therapists did not perform as well in the objective assessment made by the tool, as only 3 out of the 7 cases corroborated this improvement in their performance with the tone activity in *PreLingua*.

In general, there was a certain agreement between the tool and the therapists on when a user had improved, although there were a number of subjects with significant improvements in how well they performed different activities which were not assessed by the therapists. However, the comparability between the objective measurement provided by the tool in terms of MSE between a pattern and the subject's production cannot be considered to be complete.

8. Conclusions

The main conclusion of this work is that the education of altered voices is a difficult task that can take advantage of the use of speech technologies. A free computer-aided tool for voice therapy called *PreLingua* has been presented, which aims to train voice skills like voice activity detection, intensity, blow, tone, vocal onset, phonation time and vocalization. Inside *PreLingua*, the robustness in estimating reliable formant frequencies for all possible users and its normalization through the vocal tract length, allows to enhance the performance of the tool minimizing the inter-speaker variability of speech. This is reflected in *Articula*, a novel tool for Spanish vowel training in real time which provides a natural user interface to train vocalic articulation in voice therapy.

Promising results were obtained in a preliminary study in centers for special education during 12 weeks of therapy with *PreLingua*. The study showed the ability of an automated tool like the one proposed in this work to educate impaired users with voice alterations. These improvements were asserted by the own patients' therapists and the study performed over the results provided by the tool in terms of how well the user can use his/her voice to follow a certain pattern in intensity, blow, tone or vocalization also indicated this improvement with variable degrees of significance through many of the subjects.

The results were very encouraging to keep working in this direction, with the overall aim of providing further functionalities and robustness to the system. *PreLingua* is currently a good alternative for voice therapy in Spanish language for many speech therapists in Spain and Latin America, thanks to its distribution through the "Comunica" framework. The therapists can use this tool in different areas related to special education easing their daily work or complementing traditional techniques in voice therapy. In this area, any small contribution to the voice skills of users with voice disorders supposes an improvement in the quality of life of these individuals, enabling them to communicate more efficiently.

Acknowledgments

This work was supported under TIN-2008-06856-C05-04 from MEC of the Spanish government and Santander Bank scholarships. The authors want to acknowledge Center of Special Education "CEDESNID" in Bogotá (Colombia), and the Public School for Especial Education "Alborada" in Zaragoza (Spain), for their collaboration applying and testing *PreLingua*.

Appendix A

Tables A.1, A.2, A.3.

References

Allegro Graphic Engine. http://www.liballeg.org.

Arias, C., Estape, M., 2005. Disfonía Infantil. (Ed). Ars Medical, Barcelona, Spain.

Aronso, A., 1993. Clinical Voice Disorders, third ed. Thieme, NewYork, USA (Chapters 4, 5, 6).

CEDESNID. <http://www.cedesnid.org>.

Dr. Speech. <http://www.drspeech.com>.

Fant, G., 1960. Acoustic Theory of Speech Production. Mouton & Co., Den Hague, The Netherlands.

Games Program, Key Elemetrics. <http://www.kayelemetrics.com/ProductInfo/3950/3950.htm>.

Gurlekian, J., Elisei, N., Eleta, M., 2000. Caracterización articulatoria de los sonidos vocálicos del español de buenos aires mediate técnicas de resonancia magnética. Tech. rep., Laboratorio de Investigaciones Sensoriales. <http://www.lis.secyt.gov.ar/index.php?l=en>.

Kenneth, D., 1966. Voice therapy for children with laryngeal dysfunction. In: Proceedings of the Annual Convention of the American Speech and Hearing Association. Washington, USA.

Kirschning, I., Cole, R., 2007. Advances in Audio and Speech Signal Processing: Technologies and Applications. In: Perez-Meana, H. (Eds.). Idea Group, Hershey PA, USA. (Chapter XIV: Speech Technologies for languages therapy2).

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

La Alborada. http://centros6.pntic.mec.es/cpee.alborada.

Martínez-Celdrán, E., 1989. Fonología General y Española: Fonología Funcional. Ed. Teide, Barcelona, Spain.

Necioglu, B., Clements, M., Barnwell, T., 2000. Unsupervised estimation of the human vocal tract length over sentence level utterance. Acoust. Speech Signal Process. 3, 1319–1322.

Rabiner, L., Shafer, R., 1978. Digital Processing of Speech Signals. Prentice-Hall (Chapter 4).

Rabiner, L., Shafer, R., 2007. Introduction to Digital Speech Processing. The Essence of Knowledge, Santa Barbara CA, USA.

Rodríguez, W.-R., Lleida, E., 2009. Formant estimation in children's speech and its application for a spanish speech therapy tool. In: Proceedings of the 2009 Workshop on Speech and Language Technologies in Education (SLaTE). Wroxall Abbey Estates, United Kingdom.

Sakhnov, K., Verteletskaya, E., Simak, B., 2009. Approach for energy-based voice detector with adaptive scaling factor. IAENG Internat. J. Comput. Sci. 36 (4) (IJCS 36-4-16).

Saz, O., Yin, S., Lleida, E., Rose, R., Vaquero, C., Rodríguez, W., 2009. Tools and technologies for computer-aided speech language therapy. Speech Comm. 51 (10), 948–967.

Shahidur, M., Shimamura, T., 2005. Formant frequency estimation of high-pitched speech by homomorphic prediction. Acoustic Sci. Technol. 26 (6), 502–510.

Speech Viewer. http://www.synapseadaptive.com/edmark/prod/sv3/.

Traunmuller, H., Eriksson, A., 1997. A method of measuring formant frequencies at high fundamental frequencies. In: Proceedings of EuroSpeech'97, vol. 1, pp. 477–480.

Verhelst, W., Steenhaut, O., 1986. A new model for the short-time complex cepstrum of voiced speech. IEEE Trans. Acoust. Speech Signal Process. 34 (2), 43–51.

Vocaliza. http://www.vocaliza.es.

VoxGames. http://www.ctsinformatica.com.br.

Wakita, H., 1977. Normalization of vowels by vocal tract length and its application to vowel identification. IEEE Trans. Acoust. Speech Signal Process. ASSP-25 (2), 183–192.

Watt, D., Fabricius, A., 2002. Evaluation of a technique for improving the mapping of multiple speakers' vowel space in the f1–f2 plane. Leeds Working Papers in Linguistics and Phonetics 9, 159–173.

Please cite this article in press as: Rodríguez, W.R. et al. A prelingual tool for the education of altered voices, Speech Comm. (2011), doi:10.1016/j.specom.2011.05.006