ARTICULA - A tool for Spanish Vowel Training in Real Time

William R. Rodríguez[1], *Oscar Saz*[1] *and Eduardo Lleida*[1]

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

Abstract

This paper describes a free tool for the training of Spanish vowels called ARTICULA. This tool shows an accurate approximation in real time to the position and movements of the tongue, jaw and lips during vowel utterances independent of user's characteristics as age and gender. At the same time, the tool displays acoustic parameters as intensity, pitch, formants and spectrum, thus making ARTICULA a good alternative for vocalic articulation in voice therapy and training in Spanish language studies. The tool uses a formant normalization through the vocal tract length in order to reduce the high variability between speakers. A preliminary study in voice therapy in children with voice disorders shows the adequate biofeedback provided by the tool and, the improvement in specific vowels after ten weeks of therapy.

Index Terms: vocalic articulation, vocal tract length estimation, formant normalization, voice therapy.

1. Introduction

Vowels are mainly determined by their two lowest formants ($F1$ and $F2$) in any language, but is well known the high variability in formant values between speakers due to several factors as gender, age, or different anatomical configurations in their vocal tracts. In addition, the formant estimations are less reliable in high-pitched speech, typical in children, because the estimations vary with the continuous growing, gender, hormonal changes, and by the own speech disabilities in cases of handicapped children[1].

In logopedic area and special education cases the situation is more difficult because the few tools that works vocalic articulation (only in English and not free), show an image by plotting the $F1$ against $F2$ for a given vowel, only understood by the therapist, so there is not adequate feedback to the child.For instance, **Dr. Speech**[1] developed by Tiger DRS works real time English's vowels, using a graphic display of $F2$ against $F1$. In **Video Voice**[2] Speech Training System, the vowels appear in regions of the screen, according to $F2$ against $F1$ position. In other applications, the vocalic articulation is more "natural" but are based on videos previously recorded without considering the user's voice.

The tool proposed here allows vocalic articulation training in real time through a novel, simple and friendly user interface. Using traditional speech techniques like LPC and homomorphic analysis, and estimation of the vocal tract length, it is possible to normalize the formant frequencies in order to reduce the variability between speakers [2]. The objective in ARTICULA is to transmit to the child how to place all the organs in the vocalic productions. To reach this goal, a representation of the vocal tract is shown to the user in which these organs (tongue, jaw, lips) vary along the vocal cavity, depending on the estimated formant frequencies. This tool is oriented to the five Spanish vowels: $/a/$, $/e/$, $/i/$, $/o/$ and $/u/$.

The paper begins in Section 2 by reviewing the speech technologies inside ARTICULA. It first addresses the problem of robust formant estimation in high pitched voices, then explaining the estimations of the vocal tract length and finally the formant normalization. Section 3 describes the tool designed and how it works, Section 4 the results of preliminary study applied in children with voice disorders in special centers of education in Colombia and Spain, and finally the conclusions in Section 5.

2. Technologies Applied in ARTICULA

Traditional speech technologies like LPC analysis, homomorphic analysis and liftering, estimation of the vocal tract length, and formant normalization, are used in ARTICULA in order to reduce the formant variability inter-speaker, and to apply these in the control of ARTICULA's avatar.

2.1. 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 LPC method decreases as the fundamental frequency F0 of speech increases. Fig. 1 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 is increasing 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 $F0$. An alternative to reduce the aliasing in the autocorrelation sequence is using the homomorphic analysis as proposed in [3]. 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 $h[n]$ as in Equation (1).

$$x[n] = e[n] * h[n] \tag{1}$$

The way in which such separation is achieved is through linear

[1] http://www.drspeech.com

[2] http://www.videovoice/com/vv-frmnt.htm

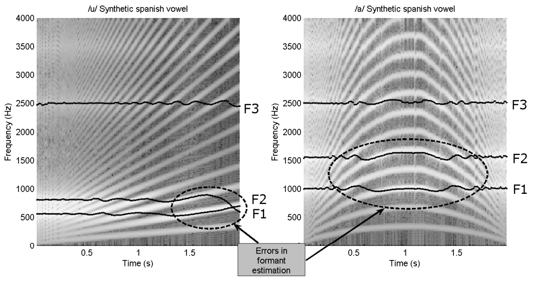


Figure 1: *High-pitch influence in formant estimation for synthetic /u/ and /a/ Spanish vowels.*

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[4], the real-domain cepstrum $c[n]$ defined by Equation (2) is used, where $X(k)$ is the N-point Fourier transform of the speech signal $x[n]$.

$$c[n] = \frac{1}{N}\sum_{k=0}^{N-1} ln\left|X(k)\right| e^{j\frac{2\pi}{N}kn}, 0 \le n \le N-1 \qquad (2)$$

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 [5] or $0.6 - 0.7T_{pitch}$ in [3]. In this work, the liftering window $w[n]$ (Equation (3)) is $0.65T_{pitch}$ and the effect of appliying $w[n]$ in the real cepstrum domain can be observed in Fig. 2.

$$w[n] = [0.65T_{pitch}, N-1-0.65T_{pitch}] \qquad (3)$$

After the liftering process, the formant frequencies $\tilde{F}_k$ without pitch influence are obtained through similar LPC method described above, with order $p = 8$ and 25 ms speech frame.

2.2. Vocal Tract Length Estimation

This section will describe a robust method to estimate the vocal tract length from formant frequencies. Modeling the vocal tract as a uniform lossless acoustic tube, its resonants frequencies given by Equation (4) are uniformly spaced, where $v = 35300$ cm/s is the speed of sound at $35\,^{\circ}\mathrm{C}$, and l is the length of the uniform tube in cm.

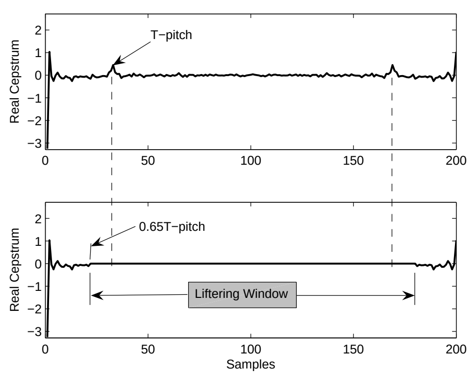


Figure 2: *Effect of liftering in the real cepstrum domain.*

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

The estimation of the length was proposed in [6], 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 Equation (5), where $D(\tilde{F}_k, (2k-1)F1)$ is a function that express the difference between the measured formant $(\tilde{F}_k)$ and the resonance of the uniform tube.

$$\varepsilon = \sum_k D(\tilde{F}_k, (2k-1)F_1) = \sum_k D(\tilde{F}_k, (2k-1)\frac{v}{4l}) \quad (5)$$

From [6], the error measure given in Equation (5) can be turned in Equation (6) using the distance function between the measured formant $\tilde{F}_k(k = 1, ..., M)$ and the odd resonances of a uniform tube, $(2k-1)F_1$.

$$\varepsilon = \sum_k \frac{(\frac{\tilde{F}_k}{2k-1} - F_1)^2}{F_1} \qquad (6)$$

Finally, minimizing Equation (6) into Equation (7) 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 Equation (8).

$$F_1 = \left(\frac{1}{M}\sum_k \left(\frac{\tilde{F}_k}{2k-1}\right)^2\right)^{1/2} \qquad (7)$$

$$VTL = \frac{v}{4F_1} \qquad (8)$$

This method was applied in a previous study[2] where based on speech records from 235 healthy children, it was possible to find a correlation between the vocal tract length and the height in children from 3 to 17 years old.

2.3. Formant Normalization

With the formant frequencies $\tilde{F}_k$ obtained in Section 2.1, we can normalize these estimations through the vocal tract length estimated in Section 2.2. The formant normalization used in this study has been proposed by [7]. 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 $\tilde{F}_k$ by the length factor, l/l_R, with l_R fixed at $17.5cm$. As shown in Equation (9), the l correspond to the vocal tract length VTL obtained from Section 2.2.

$$F_{kN} = \frac{l}{l_R}\tilde{F}_k \qquad (9)$$

A graphic comparison between unnormalized formants $\tilde{F}_k$ and normalized formants F_{kN} can be appreciated in Fig. 3. 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 its improve after normalization process (right).

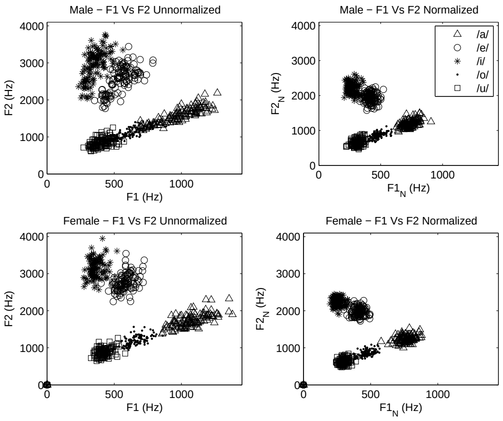


Figure 3: *Comparison between $\tilde{F}_k$ and F_{kN} for Spanish vowels from 235 healthy children -125 male (up), 110 female (down)-.*

3. ARTICULA Tool

With the robust formant estimation process (Section 2.1) and the normalization method (Section 2.3) it is possible to obtain a low variability inter-speaker in formant values. It is well known that the vocalic sounds are determined primarily by tongue position, the degree of constriction and the lip shape. From acoustical point of view the vowels can be identified by two lowest formants, $F1$ and $F2$.

ARTICULA uses the two lowest normalized formants F_{1N} and F_{2N}, in order to animate a boy or girl avatar. The tool (Fig. 4) is powered by Allegro[3] graphic engine. 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 height dimension of the tongue and $F2$ is correlated with tongue frontness [8].

As shown in Fig. 5-left, the tongue has two degrees of freedom whilst jaw has only one, whose Cartesian coordinates are described by:

$$tongue(x_t+\alpha F_{2N}, y_t+\beta F_{1N}) \qquad jaw(x_j, y_j+\gamma F_{1N}) \quad (10)$$

where x_t, y_t and x_j, y_j are the coordinates on screen position (in pixels) of the tongue and jaw respectively at rest stage, and α, β and γ are the scale factors to fit the formants frequencies units (Hz) to the screen space in pixels. In this case $\alpha = 0.022$, $\beta = 0.063$ and $\gamma = 0.03$.

The lips model (Fig. 5-right) 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 $p1, ..., p6$ are governed by the expressions:

$$p1 = (x_1 + \Delta x, y_1) \quad (11)$$

$$p2 = (x_2, y_2), \qquad p3 = (x_3, y_3), \qquad p4 = (x_4, y_4) \quad (12)$$

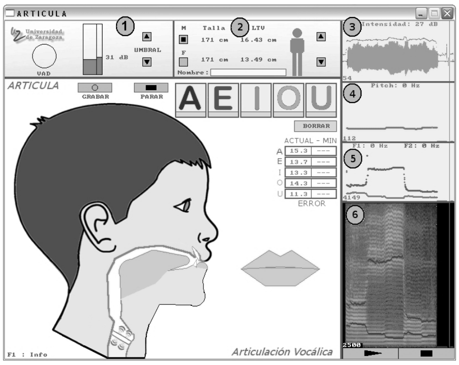
$$p5 = (x_5, y_5 + \Delta y), \qquad p6 = (x_6, y_6 + \Delta y) \quad (13)$$

where: x_i, y_i with $i = 1, ..., 6$ are the coordinates on screen position (in pixels) for each point, and Δx, Δy are the factors which moves the lips properly and are defined by:

$$\Delta x = k_1 \delta, \qquad \Delta y = 0.85\gamma F_{1N} \quad (14)$$

where δ, is the distance $\delta = \sqrt{F_{1N}^2 + F_{2N}^2}$ obtained from normalized formant frequencies, the distance δ provides a proportion of the distance between angles of the mouth. Close vowels have lower formant values (lower δ) than open vowels do (higher δ), especially in rounded vowels like /o/ and /u/. k_1, is the scale factor to fit the distance δ to 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 vertical component to the jaw. The left side of Fig. 6 shows the complete joint model, and the right side the final interface. The therapist selects the gender, height and vowel to train, and the system shows a pattern-vowel (thick line in Fig. 6-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)[9] for each vowel. The goal is 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. 4 in items 3,4,5 and 6. ARTICULA is part of "PreLingua" [10], which gathers a set of small games to train children with voice disorders, aiming to help the work in voice therapy oriented to phonation. "PreLingua" and other computer-aided tools for speech therapy withing "COMUNICA" project [11], are free and available in the project's url[4].


ARTICULA
M Talla 171 cm LTV 16.43 cm
F 171 cm 13.49 cm
Nombre:
31 dB
Intensidad: 27 dB
Pitch: 0 Hz
F1: 0 Hz F2: 0 Hz

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

Figure 4: *ARTICULA. 1-Threshold voice, 2-Gender and height selection, 3-Speech signal and Intensity outline, 4-Pitch evolution, 5-Unnormalized Formants $\tilde{F}_1$ and $\tilde{F}_2$, 6-Spectrum.*

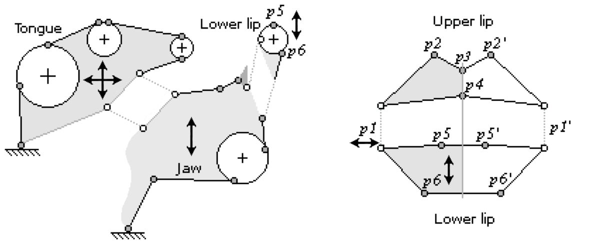


Figure 5: *ARTICULA. Tongue and jaw models (left), lips model (right).*

[3] http://www.liballeg.org Allegro

[4] http://www.vocaliza.es

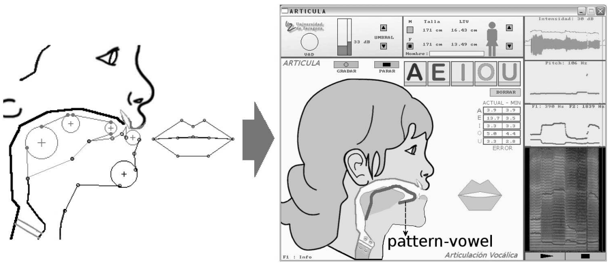


Figure 6: *ARTICULA. Joint model (left) and final interface (right).*

4. Preliminary Study and Results

In a preliminary study, ARTICULA has been tested for ten weeks in the Center of Special Education “CEDESNID”[5] in Bogotá (Colombia), and the Special Education School “Alborada”[6] in Zaragoza (Spain). Where currently is conducting a larger study with the tool “PreLingua”, testing differents activities for voice therapy like intensity, blow, and tone. The ARTICULA test was applied in children with different disabilities which affect theirs voice skills, then, the expected results vary according to child’s level of performance. To help evaluate therapy progress, ARTICULA allows to evaluate the vowels for each session. The evaluation consists in measuring the mean square error (MSE) between the pattern-vowel showed by the system and, the avatar’s tongue moved by the user’s vowel trial. 24 users (17 male and 7 female) with differents disabilities (mental and/or motor), have participated in the study and theirs MSE values were registered each week.

Table 1: *Improvement cases from out of 24.*

	/a/	/e/	/i/	/o/	/u/
Improvement %	12.5	16.6	16.6	12.5	20
Stat. Sign. %	67	70	92	84	78

Quantitative results are showed in Table 1, with the percent of the users that have improvements after ten weeks of therapy. Here $Improvement$ means a reduction in the MSE across the sessions, and session means a therapy where the five vowels utterances has been recorded by the system. From out of 24 cases, 12.5% cases shows a little improvement in the articulation production for vowels /a/ and /o/, which first formants $F1$ are higher than other vowels, highlighting the difficulty opening the mouth especially in vowel /a/ with the lowest statistical significance. 16.6%, 16.6% and 20% cases shows improvement in articulation production for vowels: /e/, /i/ and /u/ respectively, where the effort to opening mouth is less. 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.

About qualitative results, the therapists evaluated ARTICULA as easy to use, very attractive for all users and, an appropriate interface. Also have mentioned the possibility of applying ARTICULA in differents areas related to special education. Within theirs observations are significant progress in: **Cognitive area**: Memory, catch attention, concentration, and follow instructions. **Senso perceptual area**: Spatial location and visual perception. **Socialization skills**: Play in team, taking turns to play, help each other, healthy competition, auto demanding.

5. Conclusions

ARTICULA, a tool for Spanish vowel training in real time has been presented. This tool uses a natural user interface to train vocalic articulation in voice therapy. The improvement in estimations of reliable formant frequencies allow to optimize the performance of the tool minimizing the variability inter-speaker. In a preliminary study promising results has been obtained, a certain number of cases in children with voice disorders showed improvements in vowels utterances after ten weeks of therapy. The results are very encouraging to keep working in this direction as it is planned improve the functionality, robustness and start to extend this approach to other co-articulated sounds with consonants. Small contributions in voice skills in children with voice disorders, allows improving the quality of life and enabling them to communicate more efficiently.

6. Acknowledgements

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 to Center of Special Education “CEDESNID” in Bogotá (Colombia), and the School for Especial Education “Alborada” in Zaragoza (Spain), for their collaboration applying and testing the tool.

7. References

[1] A.-U. Kornilov, “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, September 2004, pp. 560–562.

[2] W.-R. Rodríguez and E. Lleida, “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, September 2009.

[3] M. Shahidur and T. Shimamura, “Formant frequency estimation of high-pitched speech by homomorphic prediction,” *Acoustic Sci. and Tech.*, vol. 26, no. 6, pp. 502–510, June 2005.

[4] R. Schafer and L. Rabiner, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[5] W. Verhelst and O. Steenhaut, “A new model for the short-time complex cepstrum of voiced speech,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 2, pp. 43–51, 1986.

[6] F.-N. Burhan, A.-C. Mark, and P.-B. Thomas, “Unsupervised estimation of the human vocal tract length over sentence level utterances,” in *Proceedings of ICASSP’00*, 2000, pp. 1319–1322.

[7] H. Wakita, “Normalization of vowels by vocal tract length and its application to vowel identification,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, no. 2, pp. 183–192, 1977.

[8] D. Watt and A. Fabricius, “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*, vol. 9, pp. 159–173, April 2002.

[9] N. E. J. Gurlekian and M. Eleta, “Caracterización articulatoria de los sonidos vocálicos del espanol de buenos aires mediate técnicas de resonancia magnética,” Laboratorio de Investigaciones Sensoriales, Tech. Rep., 2000.

[10] W.-R. Rodríguez and E. Lleida, “Prelingua - una herramienta para el desarrollo del pre-lenguaje,” in *Proceedings of V Jornadas en Tecnologías de Habla*, Bilbao, Spain, November 2008.

[11] W.-R. Rodríguez, O. Saz, E. Lleida, C. Vaquero, and A. Escartín, “Comunica - tools for speech and language therapy,” in *Proceedings of the 2008 Workshop on Children, Computer and Interaction*, Chania, Greece, October 2008.

[5] http://www.cedesnid.org

[6] http://centros6.pntic.mec.es/cpee.alborada