ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology in
Education (SLaTE2009)
Wroxall Abbey Estate, Warwickshire, UK
September 3-5, 2009

Formant Estimation in Children's Speech and its application for a Spanish Speech Therapy Tool

William R. Rodríguez, Eduardo Lleida

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

Abstract

This paper addresses the problem of how to estimate reliable formant frequencies in high-pitched speech (typical in children), and how to normalize these estimations, independent from vocal tract shape or length. The normalized formant frequencies are used to improve the performance of a Computer-Aided Speech Therapy Tool (CASTT) in Spanish. For this purpose, a study was conducted to see what is the relationship between child's height and their vocal tract length, using traditional technologies in speech processing like linear prediction LPC, homomorphic analysis and modeling of the vocal tract. Results of this study show a high correlation between child's height and their vocal tract length. The study is based on speech from 235 healthy children (110 females and 125 males) which contains Spanish vowels utterances, and enables calibration of a CASTT system for children with speech disorders.

1. Introduction

Formant frequencies are related to the articulatory act and give information about vocal tract configuration. This information is very useful in CASTT applications, which aim is to improve the quality of speech production and articulation, especially in cases when the vocalization is the main problem. Regarding to the development of CASTT, many European projects related to speech technology and speech therapy such as Orto Logo-Paedia [1], SPECO [2], ans ISAEUS [3] have been carried out during the last decade, some of them resulting in the development of software applications for speech therapy. However, there are no versions of these softwares available in Spanish, thus they cannot be used by speakers and speech therapists to train vocalization skills in Spanish.

Furthermore the formant frequencies estimation is more complex in children than in adults. Due the continuous growing, gender, hormonal changes, and others. Techniques for minimizing the high pitch influence like liftering in the cepstral domain, and the normalization of the vocal tract length are used to try to solve the problem and optimize the formant estimation in a real time application for CASTT system.

This article, which explains the work carried out, is organized as follows: Section 2 shows how the pitch affects the formant estimation. Section 3 describes how to estimate better formant values. Section 4 describes the CASTT application, and finally, Section 5 and 6 shows the discussion and conclusions respectively.

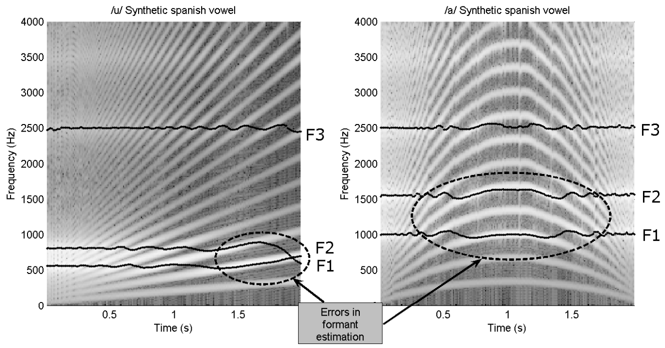

Figure 1: *Pitch influence in formant estimation*

2. Problem Identification

The formant measurement is technically difficult. The situation is less severe in male adult cases in which the fundamental frequencies (F0) are low [4]. In female and children cases F0 increase, here F0 and its harmonics could match or to be very near to the formant values affecting the real estimation. 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 of the autocorrelation sequence. In other words the accuracy of LPC method decreases as the fundamental frequency F0 of speech increases [5]. Figure 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 pitch harmonic (dashed ellipse) situation which hides the real value of the formant.

3. Formant Estimation and Normalization

Knowing the problem, the influence of the high pitch (excitation source) in formant estimation (vocal tract impulse response), we need to separate this effect in order to obtain formants "free" from F0 variations. This is possible using the homomorphic speech processing, and the estimation of the vocal tract length in order to normalize the formant frequencies constrained to children's height. This work was carried out with 235 speakers (110 females and 125 males) from 3 to 17 years old, and

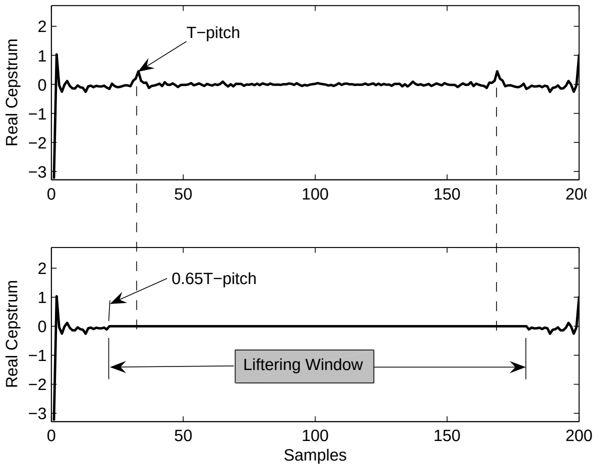

Figure 2: *Liftering in Real Cepstrum*

theirs utterances contain the five Spanish vowels /a/, /e/, /i/, /o/ and /u/.

3.1. Homomorphic Analysis

The main idea of 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 $h[n]$. So, we can write

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

The way in which such separation is achieved is through linear filtering of the inverse Fourier transform of the log spectrum of the signal[6]. The complex cepstrum is not suitable because of its high sensitivity to phase[6]. Instead of this, the real cepstrum $c[n]$ defined by equation (2) is used.

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

where $X(k)$ is the Fourier transform of the speech signal $x[n]$ using N-point. The low-time part of $c[n]$ corresponds primary to the vocal tract pulse information, while the high-time part is due primarily to the excitation. Knowing previously the value of pitch period T_{pitch} from the initial LPC analysis using autocorrelation method, it is possible lifter the cepstrum signal directly and use the new liftered signal to find the formant frequencies. A liftering window with the length of $0.5T_{pitch}$ has been proposed in [7] or $0.6 - 0.7T_{pitch}$ in [5]. Here the liftering window $w[n]$ is $0.65T_{pitch}$ and is defined in equation (3). The effect of apply $w[n]$ in the real cepstrum signal can be observed in Figure 2.

Figure 3 shows the same synthetic vowels after liftering. The effect of the pitch and its harmonics are removed and the new formant values estimated F_k are better than before.

$$w[n] = \left\{\begin{matrix} 0 & 0.65T_{pitch} \le n \le N-1-0.65T_{pitch} \\ 1 & other\ n \end{matrix}\right\} \tag{3}$$

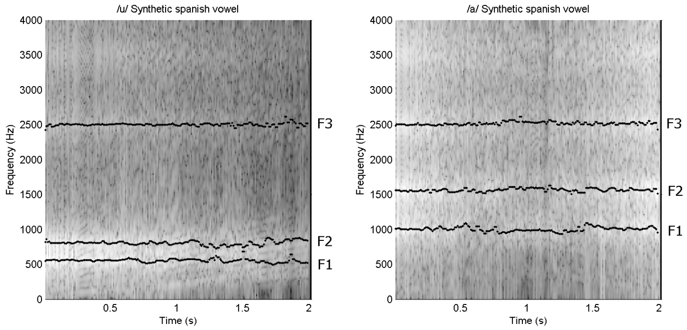

Figure 3: *Syntetic vowels after liftering*

3.2. Vocal Tract Length Estimation

As the CASTT application is applied to children, and their speech parameters (pitch and formants) change according to the growth, it is necessary to normalize the formant values in order to optimize the performance application. A reasonable indicator of the behavior of the formants in children is the vocal tract length (VTL). If we know the children's height we can infer the behavior of the formants through VTL. Modeling the vocal tract as a uniform lossless acoustic tube, closed at one end and open at the other, its resonants frequencies given by equation (4) are uniformly spaced.

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

where $v = 35300$ cm/s is the assumed speed of sound at 35 °C, and l is the length of the uniform tube in cm. The estimation of the length parameter has been proposed in [8], where the estimation can be assumed to reduce to fitting a set of known, or measured resonance frequencies of a uniform tube, which are determined solely by its length l, and given by $f_n = nv/4l$. Therefore, the problem can be approximated to minimizing

$$\varepsilon = \sum_n D(\tilde{f}_n, nf_1) = \sum_n D(\tilde{f}_n, n\frac{v}{4l}) \tag{5}$$

where $D(\tilde{f}_n, nf_1)$ is a function that express the difference between the measured resonances $(\tilde{f}_n)$ and the resonance of the uniform tube. From [8], the error measure given in equation (5) can be constructed using the distance function between the measured formant frequencies $F_k(k = 1, ..., M)$ and the odd resonances of a uniform tube, $(2k-1)f_1$. This error function is

$$\varepsilon = \sum_k \frac{(\frac{F_k}{2k-1} - f_1)^2}{f_1} \Longrightarrow \tilde{f}_1 = (\frac{1}{M}\sum_k (\frac{F_k}{2k-1})^2)^{1/2} \tag{6}$$

finally, the VTL can be obtained with the expression (7)

$$VTL = \frac{v}{4\tilde{f}_1} \tag{7}$$

Applying this approach and using the formant frequencies F_k estimated in section 3.1, the VTL estimated for 235 children are shown in figure 4. This figure shows a high correlation function between height and VTL (correlation coefficient male:0.79, female:0.62). Hence, from the measurable value of child's height we can estimate the VTL making a linear interpolation in this function in order to normalize the formant frequencies.

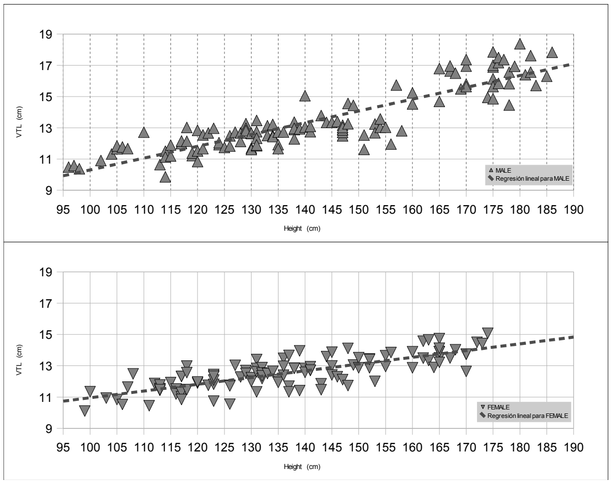

Figure 4: *VTL Estimated for 235 children (110 female, 125 males)*

3.3. Formant Normalization

With the formant frequencies F_k obtained in section (3.1) we can estimate the VTL as in section (3.2). The formant normalization used in this study has been proposed by [9]. This 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 F_k by the length factor, l/l_R, with l_R fixed at $17.5cm$. As show equation (8). The l correspond to VTL obtained from linear interpolation from figure 4.

$$F_{kN} = \frac{l}{l_R} F_k \tag{8}$$

A graphic comparison between unnormalized formants F_k and normalized formants F_{kN} can be appreciated in figure 5. This figure shows a high dispersion in unnormalized formant values F_k (left side), and how it improves after normalization. The normalized formant values F_{kN} (right side) has less dispersion than before.

4. CASTT Application

The development of Computer-Aided Speech Therapy Tools for Spanish, is currently being carried out under the framework of "Comunica"[10], which main goal is helping the daily work of speech therapists. All the tools are distributed under a freeware license for the use of all the community of speech therapists and speech therapy users who could be interested in them[1]. "Comunica" contains three applications, first "PreLingua" which works the pre-language stage and include voice activity detection, intensity, breathing and intonation control, and vocalization; "Vocaliza", which works on the phonological, semantic and syntactic levels of language, and finally "Cuéntame" which works the pragmatic level of language.

The speech processing technologies of this study are applied in $vocalization$ section of "PreLingua"[11]. "PreLingua" gathers a set of small games organized as a pyramid form and use speech processing to train children with speech development disorders in order to assist the work in speech therapy oriented to phonation. The top of pyramid is the $vocalization$ game because represents the transition from phonation, intonation and intensity control of speech, to articulation in small children. The set of vowels for every language is unique, so the strategy in the $vocalization$ game has been to make the development only for the vowels in Spanish, although expansions to any other language could be done by defining the representations of the vowels of that language to the first and second formants (F_1 and F_2) space. Spanish language contains five vowels ($/a/$, $/e/$, $/i/$, $/o/$ and $/u/$ in their SAMPA notation) whose representation in the space of the formant frequencies F_1 and F_2 is a triangle, like in figure 5.

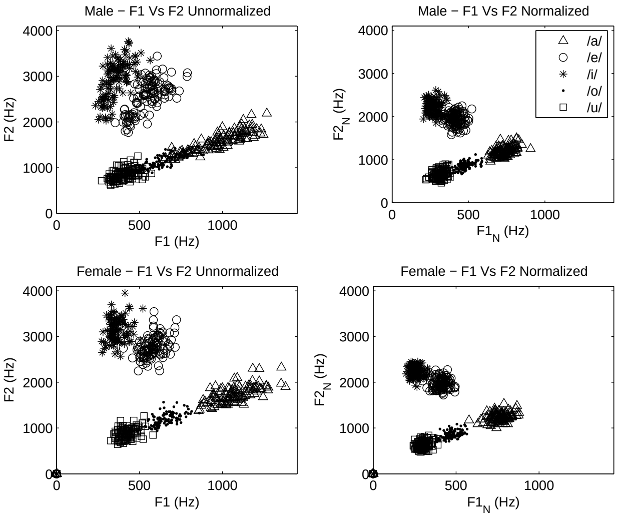

Figure 5: *Comparison between F_k and F_{kN}*

At first, the game asks the child's gender and height, then based on interpolation data in figure 4 the system set up the "standard" F_{1N} Vs F_{2N} region for each vowel according to formants for a healthy child with same height. After that, the idea is encouraging the user to utter each vowel and the system draws a dot according to F_{1N} Vs F_{2N} estimated. If the pronun-

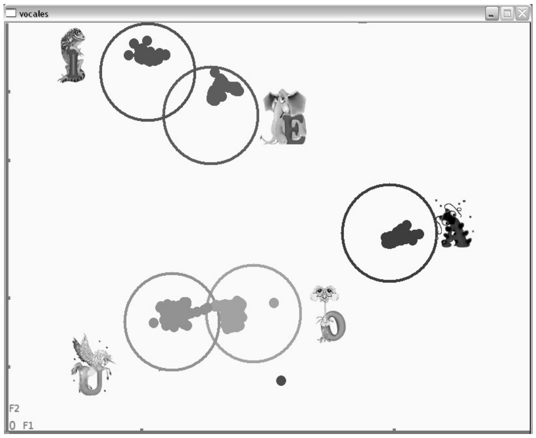

Figure 6: *Vocalization game*

[1] URL http://www.vocaliza.es

ciation is close to the correct form the system draw the dot with the same color of vowel and will animate a cartoon, otherwise, the dot will be draw with different colors and the cartoon won't animate. At this manner, the child can see in real time the effects of his/her own utterance. The Figure 6 shows the result of pronounce the five vowels of correct form.

5. Discussion

The discussion of this work was focused on whether it was possible or not to normalize the formant frequencies in high-pitched speech typical in children, using any parameter or reference from child directly. This was possible by means of application of different technologies, like liftering in the cepstral domain (section 3.1) and vocal tract length estimation (section 3.2). Based on speech records from 235 healthy children, it was possible find a "theoric" relationship between the vocal tract length and the height in children from 3 to 17 years old. After liftering in cepstral domain the formant frequencies obtained F_k are normalized in F_{kN} using equation (8)(section 3.3), where l_R is fixed at $17.5cm$ and l is the VTL obtained by linear interpolation from data in figure 4 using the height in cm asked previously for the system.

Figure 5 shows the original estimations F_k (left side) and the results after normalization F_{kN} (rigth side) for 235 children utterances. In males cases the F_k estimations (left-up) shows a big dispersion in semi-open vowels ($/e/$ and $/i/$), but the estimations are improved after the normalization process where the dispersion in F_{kN} (right-up) is significantly lower. Furthermore, another improvement can be seen comparing the F_{kN} results for male and female normalized cases (right-side), where is possible to use the same template of vocal triangle independent of gender. Hence, this work has proven that the use of cepstral liftering, vocal tract length estimation and formant frequencies normalization from body height, is possible and improve the estimation of the formant frequencies in children's speech for CASTT applications.

6. Conclusions

As a conclusion of this work, the normalization of formant frequencies in children's speech is possible by means of technologies like homomorphic analyze and the estimation of vocal tract length from the body height of child. The improvement in the formant frequencies estimation allow to optimize the performance of Computer-Aided Speech Therapy Tools for Spanish like "PreLingua". This tool are in use at a special education school $Alborada$ in Zaragoza, where have been very successful among the children and speech therapists, the evaluations by the group of speech therapists shows that is a useful tool for the training of children with speech disorders in the pre-language stage.

The therapists also evaluate positively the easiness of use of the tool. The results are very encouraging to keep working in this direction as it is planned improve the functionality and robustness. Further work in this area might include a better estimations of articulation parameters in children and its applications in systems based on automatic speech recognition ASR. The adequate development of pre-Language stage in children improves the quality of life of individuals with speech disorders and enable them to use computers by means of multimedia applications.

7. 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 CEIP "Rio Ebro" school for their collaboration in data acquisition and speech therapist from the CPEE "Alborada" for their ongoing collaboration and test in this work.

8. References

[1] A.-M. Oester and D. House and A. Protopapas and A. Hatzis, "Presentation of a new EU project for speech therapy: OLP (Ortho-Logo-Paedia)", Proceedings of the XV Swedish Phonetics Conference (Fonetik 2002), May, 2002.

[2] K. Vicsi and P. Roach and A. Oester and Z. Kacic and P. Barczikay and I. Sinka, "SPECO: A Multimedia Multilingual Teaching and Training System for Speech Handicapped Children", Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech-Interspeech), September, 1999.

[3] R. García-Gómez and R. López-Barquilla and others, "SPEECH TRAINING FOR DEAF AND HEARING IMPAIRED PEOPLE: ISAEUS Consortium", Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech-Interspeech), September, 1999.

[4] Hartmüt Traunmuller and Anders Eriksson, "A Method of Measuring Formant Freqiencies at High Fundamental Frequencies", Eurospeech, 1997.

[5] M. Shahidur Rahman and Tetsuya Shimamura, "Formant frequency estimation of high-pitched speech by homomorphic prediction", Acoustic Sci. and Tech., June, 2005.

[6] L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signals", Chapter 7, 1978.

[7] 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., 34,43-51, 1986.

[8] Burhan F. Necioglu, Mark A. Clements and Thomas P. Barnwell, "Unsupervised Estimation of the Human Vocal Tract Length Over Sentence Level Utterances", Acoustics Speech and Signal Processsing., ICASP00, 2000.

[9] HISASHI WAKITA, "Normalization of Vowels by Vocal Tract Length and its Application to Vowel Identification", IEEE Transactions on Acoustics, Speech and Signal Processing., VOL. ASSP-25, NO. 2, April 1977.

[10] William R. Rodríguez, Oscar Saz, Eduardo Lleida, Carlos Vaquero and Antonio Escartin, "COMUNICA - Tools for Speech and Language Therapy", Workshop on Child Computer and Interaction., ICMI08 post-conference workshop, Chania, Crete, Greece, October 2008.

[11] William R. Rodríguez, Eduardo Lleida, "PRELINGUA - Una Herramienta para el Desarrollo del Pre-Lenguaje", V Jormadas en Tecnologa del Habla, Bilbao - Spain, November 2008.