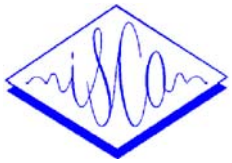

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SPEAKER NORMALIZATION FOR AUDIO-VISUAL ARTICULATION TRAINING

Marcel Ogner and Zdravko Kacic

University of Maribor, Faculty of Electrical Engineering and Computer Science
WWW: http://www.dsplab.uni-mb.si

10.21437/Eurospeech.1999-149

1. ABSTRACT

The paper describes formant based speaker normalization method suitable for speech visualization and articulation training systems. The method estimates the error function obtained from speaker formant characteristics for a given vowel. Estimated error function gives information for critical band filter shifting on mel-warped frequency scale. The paper also describes accurate technique for formant tracking.

keywords: normalization, formant, critical bands, frequency warping, polynomial interpolation

2. INTRODUCTION

Usually the problem of speaker normalization is considered as technique used to increase recognition accuracy in speaker-independent speech recognition task. In this paper we discuss the use of speech normalization algorithms in a teaching and training system for speech handicapped children. The task of the speaker normalization is to provide uniform visual representation of the speech characteristics to enable efficient training of correct pronunciation of vowels and fricatives for hearing-impaired children.

The work is part of the INCO-COPERNICUS project Speco (A Multilingual teaching and training system for speech handicapped children) with the goal to extend the applicability of the system to wider children age range -- also to the group of children older than 10 years for which the pupils voices start to differ significantly in fundamental frequency and other characteristics. In this cases the efficient use of the system is not possible without some sort of speaker normalization techniques that will enable suitable presentation of speech characteristics. The visual representation of speech is done by using the critical-band spectrum analysis. The normalization techniques used must encounter requirements of such presentation in combination with graphic images used to motivate the user. It should also encounter the fact that visual presentation of speech should give the possibility to the user to further process speech on the basis of her or his visual perception, as there usually is no auditory feedback available. This sets specific requirements for speaker normalization techniques implemented and also the need of using other assessment methods than the speech recognition error rate that is frequently used it the case of speech recognition. It should be also noted that the normalization technique should preserve all the important characteristics of the pronunciation as it will often be very pore. The child had to have possibility to observe the speech picture which should help him to develop better pronunciation following also the guidelines of the therapist

Since normalization should preserve all important information of bad articulation, all we can afford is normalization due to variations in different vocal tract shape (length) among different speakers of different ages. To derive vocal tract system characteristics, frequently used pitch synchronous formant parameter estimation will be described. Determination of the formant location require identification of the starting point of the closed phase within each pitch period. In section three, the algorithm for detecting such instants in speech signal is discussed. In section five we present an algorithm for frequency warping, where error function is extracted from formant patterns between two speakers, speech handicapped and reference speaker. Using these error function critical band filters are shifted along frequency axis in such a way that in turn results in more uniform spectral representation of correctly pronounced vowels.

3. LOCATING THE BEGINNING OF GLOTTAL CLOSED PHASE IN THE SPEECH SEGMENT

Closed phase analysis of voiced speech is based on the assumption that instants in the signal can be located that correspond to periods in which vocal folds are closed, means no excitation to the vocal tract. The algorithm which has been already well described in [3] bases on the assumption that vocal tract including glottal waveform has minimum phase characteristics [2]. The algorithm is depicted as block diagram in figure 1. The vocal tract can be modeled as

$$H(z) = \frac{G(1 - az^{-1})}{1 - \sum_{i=1}^{p} a_i z^{-i}} \qquad (1)$$

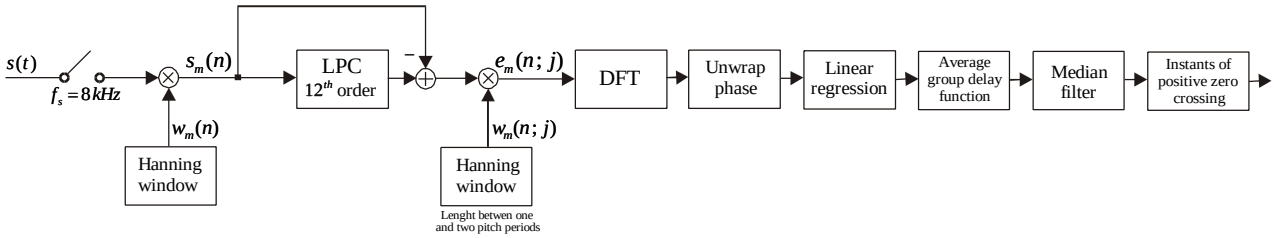

Figure 1. System for detecting instants of glottal closure

where *G* is gain factor and a is lips radiation parameter. This is a minimum phase system if a is slightly grater then 1. Equation (1) follows to difference equation

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + [Gu(n) - Gau(n-1)] \tag{2}$$

where *u(n)* denotes glottal volume velocity. Using LPC inverse filtering we can get some kind of estimation of the second part of the right side of equation (2). Note that linear prediction inverse filtering will preserve all minimum phase characteristics in the residual. The minimum phase signal is the one which, for a given magnitude spectra, has a minimum temporal delay of each frequency component in the spectrum. For a minimum phase signal starting at the origin of the analysis window, the average group delay is zero, while shifted version of such signal will have average group delay proportional to the shift. In order to derive average group delay as a function of time, also called phase slope function, we compute unwrapped phase spectrum of LPC residual for each inner window shift.

$$\arg\left\{ E_m(k\,;\, j) = \sum_{n=0}^{N-1} e_m(n) w_m(n - d_s\ j)\, e^{-jkn\frac{2p}{N}} \right\} \tag{3}$$

$$j = 0,1,.....,M - \frac{N}{2} - 1$$

$$n = 0,1,.....,N-1$$

Here d_s is window shift and is equal 1 sample, M is frame size; M=256, N is inner window size, chosen between one and two pitch periods, so it is assured that at each instant average group delay is dictated by at least one excitation peak. Slope of the linear regression fit for the resulting phase spectrum is taken as average group delay. Peaks in the residual signal (figure 2.a) coincide with the major excitation to the vocal tract which in addition correspond to the glottal closure. At instants where time origin of the analysis window coincide with the start of the minimum phase signal, the average group delay will be zero. When two peaks occurred within the window, average group delay is determined by the position of higher amplitude peak. For that reason, it is better to place the time origin at the centre of analysis window because of window tapering at the edges. Note, that phase slope function computed at each sampling instant, specially for unvoiced speech, will show some fluctuations that may results in spurious zero crossing. To avoid such problem we have used smoothing median filter which preserve sharp edges and does not shifts the signal in time domain.

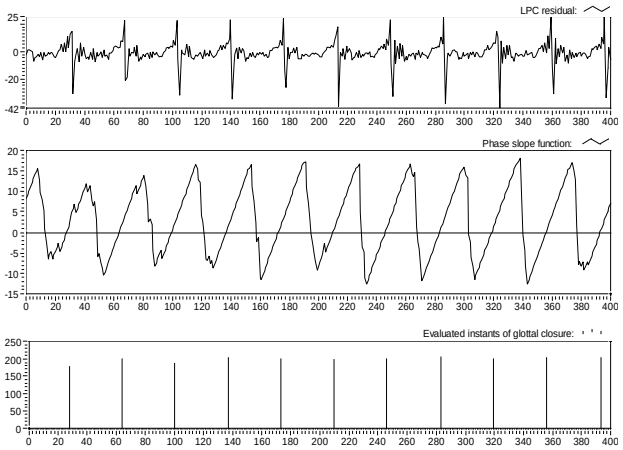

Figure 2. (a) Linear prediction residual of the speech segment for female vowel /a/; sampling frequency 8 kHz, (b) Phase slope function without smoothing; window length 64 samples, (c) Instants of glottal closure

4. FORMANT FREQUENCIES MEASUREMENT

One major source of inter-speaker variability in visual presentation of correctly pronounced voiced sounds is the variations of vocal tract shape. It is difficult to derive information about vocal tract characteristics from the speech signal directly, therefore we use formant parameters. In this paper we discuss the formant frequencies only, as they provide enough information for what critical band filter shifting technique needs. During the closed glottal cycle the short speech segments within pitch period can be precisely modeled by a p-order all-pole system. In that phase, the speech signal consist only of the free resonances of the vocal tract system. A frequently used technique for formant location involves solving for the roots of linear prediction inverse filter polynomial A(z) obtained for the above mentioned speech segments. For a given complex root pair, close enough to the unit circle, $z = r_0 e^{\pm j_0}$ the formant frequency is deduced by the angle of the complex pair

$$f = \frac{F_s}{2p} j_0 \qquad (4)$$

It is important to note that closed phase can be very short. The postexcitation speech segments were taken between 3rd sample after detected closed phase and 20th sample before next detection occur. Because of short analysis segments covariance method was used to obtain LPC parameters as it does not require analysis window. Estimated formant frequencies for each pitch period were then averaged over the 20 ms frame.

5. FILTER BANK POSITIONING

In this section we describe a speaker normalization technique based on spectral shifts in the auditory filter domain. Because the final goal is visualization of the speech production rather then automatic speech recognition we permit to ourselves somewhat more freedom. Common way to extract acoustic features from speech is to obtain the smoothed estimate of the formant envelope. Further improvement can be obtained by mapping the real frequency scale (Hz) to perceived frequency scale (mels) or even more commonly computing equal-loudness weighted total energy only in critical bands around mel frequencies using critical band filters. Visualizing these features (spectrogram, cohleogram) will poses some variability, since different speakers have different formant frequencies for the same vowel, even if it is excellent pronounced. A main source for this variability among different speakers is due to the differences in vocal-tract lengths. Conventional speaker normalization techniques use parametric approach and attempt to estimate constant scale factor between different speaker populations. In present study we use formant based approach. Let we say that reference speaker has formant pattern Fr={f_1,f_2,f_3,...,f_N} for a given vowel. For reference speaker we adopt usual mel critical band distribution, shown in picture 3, where each formant falls in certain weighted region of particular critical band filter. The patient using this articulation training system will produce somewhat different formant pattern for the same vowel. Moving his/her formants along mel-axis (figure 3) gives us N points which coincide with formants of the reference speaker according to filter index and its weighting on mel scale. But we still dont know how to shift critical band filter centered at mel frequencies to achieve formant matching. We solve this problem by polynomial fit. The model for polynomial fit is

$$y_i = \sum_{j=0}^{k-1} b_j x_i^j = b_0 + b_1 x_i + b_2 x_i^2 + \ldots + b_{k-1} x_i^{k-1}$$
$$i = 0,1,2,\ldots..N-1 \;;\; k < N \qquad (5)$$

The fiting problem reduces the problem of finding coefficients B=[b_1, b_2,..., b_{k-1}] that minimizes the difference between the observed data y_i and the predicted value. We use the least Chi-square plane method to obtain coefficients in (5), that is, finding the solution, B, which minimizes the quantity

$$c^2 = \sum_{i=0}^{N-1} \left(\frac{y_i - \sum_{j=0}^{k-1} b_j x_i^j}{s_i} \right)^2 = \left| HB - Y \right|^2 \qquad (6)$$

where H is observation matrix

$$H = \begin{vmatrix} 1 & f_0 & f_0^2 & . & . & f_0^{k-1} \\ 1 & f_1 & f_1^2 & . & . & f_1^{k-1} \\ 1 & . & . & . & . & . \\ 1 & . & . & . & . & . \\ 1 & . & . & . & . & . \\ 1 & f_{N-1} & f_{N-1}^2 & . & . & f_{N-1}^{k-1} \end{vmatrix} \qquad (7)$$

If the formants are independent and normally distributed with constant variance, $s_i = s$ the preceding equation is also the least square estimation. One way to minimize χ to set the partial derivatives of χ to zero with respect to b_1, b_2,..., b_{k-1}, which leads to matrix notation

$$H^T HB = H^T Y \qquad (8)$$

Equation (8) can be solved using LU or Cholesky factorization algorithm. Subtracting modeled function from reference gives us an error function which determines necesery shift value at each frequency.

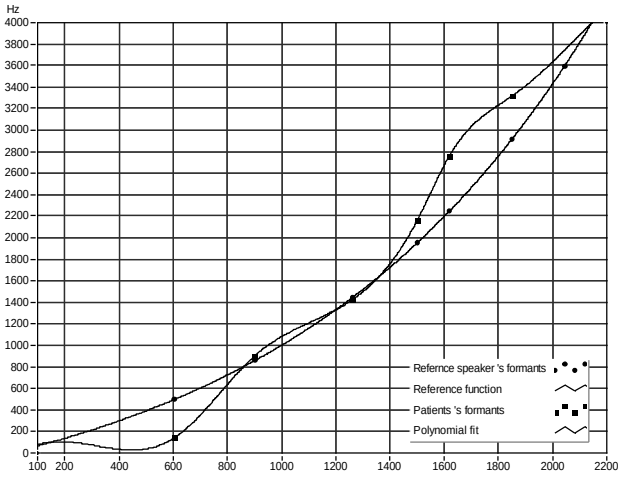

Figure 3. Formant frequencies presented in mel/Hz plane. Obtained error function determines necessary shift value at each frequency.

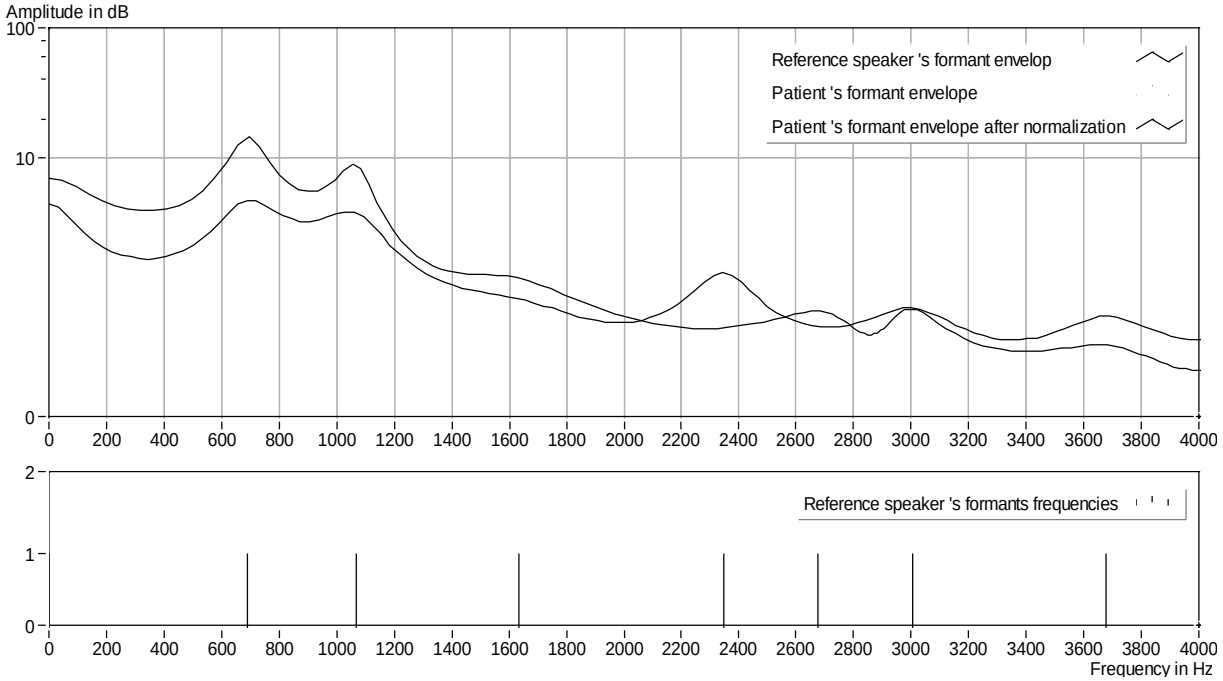

Figure 4. (a) Postexcitation formant envelope for vowel /a/ of two speakers: female reference speaker, male test speaker. (b) Formant frequencies of reference speaker.

6. RESULTS AND CONCLUSION

Figure 4 shows formant envelope of reference male speaker obtained by speech modeling in glottal closed phase. Also formant envelopes for female speaker are shown before and after the normalization takes place. Normalization has been done by shifting each spectral component according to the error function discussed in section 5. It is easy to see that formant peaks after normalization coincide with peaks of reference speaker.

The methods used has proven to be efficient in pitch-synchronous formant analysis. The speaker normalization was done using the filter positioning methods suitable for use in speech visualization and articulations training systems.

To further elaborate and estimate the methods used, experiments on larger speech database speech and hearing handicapped children voices will be performed. To test the efficiency of the method in real environment the integration into the speech corrector system (developed in the framework of the SPECO project) is foreseen.

7. REFERENCES

[1] J. D. Markel and A. H. Gray, *Liner Prediction of speech*. New York: Springer Verlag, 1976

[2] B. Yegnanarayana, Raymond N. J. Veldhuis, *Extraction of Vocal-Tract System Characteristics from Speech Signals*, IEEE Trans. on Speech and Audio Proc., VOL. 6 NO. 4, July 1998

[3] R. Smits, B. Yegnanarayana, *Determination of Instants of Significant Excitation in Speech Using Group Delay function*, IEEE Trans. on Speech and Audio Proc., VOL. 3 NO. 5, July 1995

[4] D. J. Wong, J.D. Markel and A.H. Gray, *Least squares glottal inverse filtering from the acoustic speech wave*, IEEE Trans. Acoust., Speech Signal Processing, VOL. ASSP-27, NO. 8, 1979