Monitoring of the Natural Voice Variations in Open and Closed Phases with Frequency Warped ARMA Modeling

Pedro J. Quintana-Morales, Juan L. Navarro-Mesa, Antonio G. Ravelo-García & Fernando D. Lorenzo-García

Departamento de Señales y Comunicaciones
Universidad de Las Palmas de Gran Canaria – Spain
pquintana@dsc.ulpgc.es, jnavarro@dsc.ulpgc.es

10.21437/Interspeech.2006-626

Abstract

The objective of this paper is to propose the use of a speech model with psychoacoustical information to distinguish between the open and closed phases of the vocal folds, in order to monitor formants as phonetic speech characteristics. Taking a frequency warped ARMA model for each one of the phases, the aim is to integrate the information of various consecutive periods in which the poles and the zeros that form the vocal tract can be considered common due to its slow variation. To analyze the capacity of phonetic monitoring, different phonetic voice transitions of speech registers are used from a database that provides information on glottal closings. First, we show the dependency with the warping factor. Then, the consistency and reliability of the method is demonstrated, as well as its better performance compared to no warped ones. And finally, we will see the well behavior against moderate noise.

Index Terms: ARMA modeling, frequency warped, voiced speech analysis, glottal phases, formant monitoring

1. Introduction

The modeling of speech signal which takes into account the variation of the excitation in the voiced speech production system, when the vocal folds open and close periodically [1], can be very useful in applications where we seek an efficient monitoring of the speech features. Also, the inclusion of an auditory model has revealed an important issue for applications in speech recognition [2] and coding [3].

The classical estimation methods usually work over local-stationary equally spaced frames during 30 msec approximately and include some pitch periods. However, the consideration of the different characteristics that we can observe in both closed and open phases - with stable features for the closed one since the system is composed of vocal tract only and with variable features for the open one because the system is composed of the vocal tract, the trachea and lung - shed doubt on the value usefulness of the average characteristic of the classical results [1]. The need to carry out monitoring of those natural variations of the voiced speech suggests the use of a model for each phase and a synchronous analysis with the instants of glottal closure.

Some studies, like [1] and [4], have proposed ARMA models per phase with special care given to the small analysis frame length due to the possible inconsistency in the results. If we look at slow variations of the speech production system characteristics, the estimations could be improved using the phases of several consecutive periods jointly. Therefore, a least-squares solution with a rough average of the covariance matrices concerning those periods was proposed in [1]. Also, in [4], we proposed an ARMA model formulation to directly integrate the phase information of consecutive periods. We found a least-squares solution to be suitable for estimating the common pole and the particular or common zero structure of the periods and it outperformed other classical ones based on averaging sample covariance matrices.

We have also introduced the auditory perception point of view into common pole and particular zeros ARMA model of consecutive periods [5]. This was achieved by applying a warping function that controls the frequency resolution in a psycho-acoustical manner. The analysis was based on the idea of using first-order all-pass dispersive sections instead of delay units in the difference equations of an ARMA model. This notion also had been introduced for warped linear predictive coding [3] in frame-by-frame analysis.

Now, continuing our work, we propose to develop a pitch-synchronous frequency warped common-pole common-zero (WCPCZ) model associated to several adjacent periods to minimize the reconstruction error. We study the speech natural phonetics variations across the tracking of the formants performance. We show the dependence on the warping factor and display the consistency of the estimations in the sense of regularity in the estimated formants dynamic and their transitions. We also study the reliability in the sense of the estimated formants accuracy. The behaviour of the new method is tested in relation to non warped ones. We use different phonetics transitions belonging to the database *Keele* and their corresponding glottal closure instants.

In the next section we review the warped method which performs over one period. In section 3 we expose the new frequency warped ARMA model with common poles and common zeros to several periods. In section 4 we carry out the experiments and display the results. And in section 5 we write the conclusions.

2. Frequency warped ARMA model over one period

Let *y(n,k)* be the signal associated to a given phase (open or closed) within the *n-th* period. The expression for the pole-zero, ARMA process, model is as follows

$$y(n,k) = -\sum_{i=1}^{p} a_i^n y(n,k-i) + \sum_{i=0}^{q} b_i^n u(n,k-i) \qquad (1)$$

where *u(n,k)* is the excitation signal within the *n-th* period of any phase, *{k=0,...,N_n-1}*, N_n is the phase length and *{a_i^n, i=1,...,p}* and *{b_i^n, i=0,...,q}* are the AR and MA coefficients of orders *(p,q)*, respectively. In *Z* domain, the expression (1) transforms to

$$Y_n(z) = -\sum_{i=1}^{p} a_i^n z^{-i} Y_n(z) + \sum_{i=0}^{q} b_i^n z^{-i} U_n(z) \quad (2)$$

Frequency warped ARMA model is obtained by replacing the units delay z^{-1} with all-pass first-order sections like this:

$$D(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \quad (3)$$

where, –1<λ<1, is the warping factor. Then, we obtain

$$Y_n(z) = -\sum_{i=1}^{p} a_i^n D^i(z) Y_n(z) + \sum_{i=0}^{q} b_i^n D^i(z) U_n(z) \quad (4)$$

where $D^i(z)$ is a generalized delay operator in the *Z* domain. In the time domain, this operator over a given signal *x(n)* can be defined by

$$d_i[x(n)] = \underbrace{d(n) * d(n) * \cdots * d(n)}_{i-times} * x(n) \quad (5)$$

Transforming (4) to the time domain and applying (5) we can obtain the extended expression of (1) by

$$y(n,k) = -\sum_{i=1}^{p} a_i^n d_i(y(n,k)) + \sum_{i=0}^{q} b_i^n d_i(u(n,k)) \quad (6)$$

Let's now define the reconstruction error signal as:

$$e(n,k) = y(n,k) + \sum_{i=1}^{p} a_i^n d_i(y(n,k)) - \sum_{i=0}^{q} b_i^n d_i(u(n,k)) \quad (7)$$

Using a matrix notation and assuming that *u(n,k)* is known or appropriately estimated, the error associated to all instants of the given phase corresponds to

$$\underline{e_n} = \underline{y_n} - \left[\underline{\underline{Y_n}} \quad \underline{\underline{U_n}}\right]\underline{h_n} = \underline{y_n} - \underline{\underline{H_n}}\,\underline{h_n} \quad (8)$$

$$\underline{e_n} = \left[e(n,0), \cdots, e(n, N_n + N'_n - 1)\right]^T$$

$$\underline{y_n} = \left[y(n,0), \cdots, y(n, N_n - 1), 0, \cdots, 0\right]^T, (N_n + N'_n) \times 1$$

$$\underline{h_n} = \left[a_1^n \ \cdots \ a_p^n \ \ b_0^n \ \cdots \ b_q^n\right]^T$$

$$\underline{\underline{Y_n}} = \begin{bmatrix} d_1(y(n,0)) & d_2(y(n,0)) & \cdots & d_p(y(n,0)) \\ d_1(y(n,1)) & d_2(y(n,1)) & \cdots & d_p(y(n,1)) \\ \vdots & \vdots & \ddots & \vdots \\ d_1(y(n,N_n-1)) & d_2(y(n,N_n-1)) & \cdots & d_p(y(n,N_n-1)) \\ \vdots & \vdots & \ddots & \vdots \\ d_1(y(n,N_n+N'_n-1)) & d_2(y(n,N_n+N'_n-1)) & \cdots & d_p(y(n,N_n+N'_n-1)) \end{bmatrix}$$

$$\underline{\underline{U_n}} = \begin{bmatrix} d_0(u(n,0)) & d_1(u(n,0)) & \cdots & d_q(u(n,0)) \\ d_0(u(n,1)) & d_1(u(n,1)) & \cdots & d_q(u(n,1)) \\ \vdots & \vdots & \ddots & \vdots \\ d_0(u(n,N_n-1)) & d_1(u(n,N_n-1)) & \cdots & d_q(u(n,N_n-1)) \\ \vdots & \vdots & \ddots & \vdots \\ d_0(u(n,N_n+N'_n-1)) & d_1(u(n,N_n+N'_n-1)) & \cdots & d_q(u(n,N_n+N'_n-1)) \end{bmatrix}$$

where $\underline{e_n}$, $\underline{y_n}$ and $\underline{h_n}$ are the error, signal and coefficients vectors respectively, $\underline{\underline{Y_n}}$ and $\underline{\underline{U_n}}$ are the signal and excitation matrices respectively and N'_n is an extension where there are non-zeros error terms.

3. Frequency warped common-pole and common-zero over M periods

In this section we make use of the fact that the natural variations in the characteristics of the vocal-tract system are slow with respect to the pitch period. We can assume that for each phase and for some, *M,* consecutive periods of speech, its zero structure (antiresonants) and its pole structure (resonants), keep constant from period to period. In this case [4], we can redefine equation (8) in order to make an estimation of the *n-th* period coefficients over those periods.

$$\underline{e_{nM}} = \underline{y_{nM}} - \underline{\underline{H_{nM}}}\,\underline{h_{nM}} \quad (9)$$

$$\underline{e_{nM}} = \left[\underline{e_{n+0}}^T, \cdots, \underline{e_{n+M-1}}^T\right]^T, (MxN_{nM})x1$$

$$\underline{y_{nM}} = \left[\underline{y_{n+0}}^T, \cdots, \underline{y_{n+M-1}}^T\right]^T, (MxN_{nM})x1$$

$$\underline{h_{nM}} = \left[\underline{a}^T, \underline{b}^T\right]^T, (p+(q+1))x1$$

$$\underline{a} = \left[a_1^n, \cdots, a_p^n\right]^T, \underline{b} = \left[b_0^n, \cdots, b_q^n\right]^T$$

$$\underline{\underline{H_{nM}}} = \begin{bmatrix} \underline{\underline{Y_{n+0}}} & \underline{\underline{U_{n+0}}} \\ \underline{\underline{Y_{n+1}}} & \underline{\underline{U_{n+1}}} \\ \vdots & \vdots \\ \underline{\underline{Y_{n+M-1}}} & \underline{\underline{U_{n+M-1}}} \end{bmatrix}, (MxN_{nM})x(p+(q+1))$$

where the error vectors $\underline{e_{n+j}}$ and signal vectors $\underline{y_{n+j}}$ are similar to those of equation (8), $\underline{\underline{Y_{n+j}}}$ y $\underline{\underline{U_{n+j}}}$ are the signal and the excitation matrices corresponding to the given phase of the *(n+j)-th* period, *{j=0,...,M-1}* and $N_{nM} = max\{[N_{n+0}, \ldots, N_{n+M}]\}$.

The coefficient vectors $\underline{h_{nM}}$ continue having *p+(q+1)* elements obviously. The first ones, *{a_i^n}*, correspond to the common pole structure and the remainder ones, *{b_i^n}*, correspond to the common zero structure.

We use the cost function defined as the square sum of the reconstruction error for time index *k* of the signal of a given phase within the *M* consecutive periods starting from the *n-th* one.

$$C_M(n) = \sum_{j=0}^{M-1} \sum_{k=0}^{L} e^2(n+j,k) \quad (10)$$

The coefficients $\underline{h_{nM}}$ that minimize $C_M(n)$ in (9) using the least-squares method can be represented, therefore, in vector form as

$$\underline{h_{nM}} = (\underline{\underline{H_{nM}^T}}\,\underline{\underline{H_{nM}}})^{-1}\,\underline{\underline{H_{nM}^T}}\,\underline{y_{nM}} \quad (11)$$

being the solution in (8) a particular case, $h_n = h_{n0}$.

This leads us to the least squares solution of the frequency warped common-pole common-zero (WCPCZ) model over M periods.

On the other hand, we have the non-warped multicycle alternatives solutions. One is the extended common-pole common-zero model (ECPCZ) [4] - the non warped version of WCPCZ, obtained from (1), not from (6). Another is a classical

one, given by a covariance method, averaging over *M* periods [1], which we call MCC and it is obtained by resolution of:

$$\left(\sum_{n=0}^{M-1}\underline{\underline{C}}_n\right)\underline{a}=\left(\sum_{n=0}^{M-1}\underline{c}_n\right) \qquad (12)$$

where $\underline{\underline{C}}_n$ and $\underline{c}_n$ are the *n-th* period covariance matrix and vector, respectively.

4. Experiments and results

Our goal with the following experiments was to analyze the new proposed method, WCPCZ, for tracking the speech phonetic features, like the formants. Therefore we used different phonetic transitions uttered by men and women. The study is centered on the consistency and trustworthiness with regard to spectrogram. We show the behavior with the warping factor and compare the new method to the alternatives ones, ECPCZ and MCC. Also, we use them in noisy environments to establish their robustness.

We used some voice recordings stored in the database *Keele* to carry out the experiments. The database contains 5 male and 5 female recordings (each one is about 40 seconds long) with their corresponding laryngogram. The sampling frequency f_s is *20 KHz*. Prior to the experiments we used the laryngogram signals to mark the correct glottal closure instants (GCI) (29292 were obtained), voiced/unvoiced intervals and pitch. The open and the closed phases are extracted from the GCI. In each period, the closed and opened phases represent *40%* and *48%* of each period length, respectively. As suggested in [1] the open phase ends an arbitrary (*12%* is our compromise value) instant before the excitation. In all experiments M is set to 3. Also, the *(q,p)* order is set to *(11,12)*. This seems a good choice if we consider that in *10 kHz* there are about 8 formants.

The formants are obtained from the roots of the AR part in each ARMA model, whenever they are higher than *200 Hz*, or smaller than *9.8 kHz* and their modules are larger than *0.8*. We represent the closed phase formants by symbol *"*"* and the open phase ones by symbol *"."*. We display them in the same graphic to check the spectral tracking of voice segment and observe the different characteristics of each phase. We also show the speech wave to notice the transitions and the spectrogram to see the formant way clearly.

In figure 1.a we show a voice segment /aveller/ uttered by a man and composed of vocal-consonant and consonant-vocal transitions. In this figure we can check as the WCPCZ method depends on the warping factor to track the formants. We have tried some values and put the most representative ones: *0.25* in fig. 1.c, *0.45* in fig. 1.d and *0.65* in fig. 1.e. We can see that the higher the warping factor, the better the resolution (more important perceptually) in lower frequencies (under turning point [5]) leaving few formants to track the higher frequencies. A warping factor of *0.45* is a good compromise to follow the most important phonetic characteristics for the hearing (lower band) and at the same time, to be able to track higher frequencies features (good for intelligibility). Hereafter we'll use this value as warping factor.

In the figure 2.a we display the vocal-consonant transition /an/ pronounced by a woman. In the figure 2.c we can see the variations of the estimated formants with the proposed method WCPCZ.

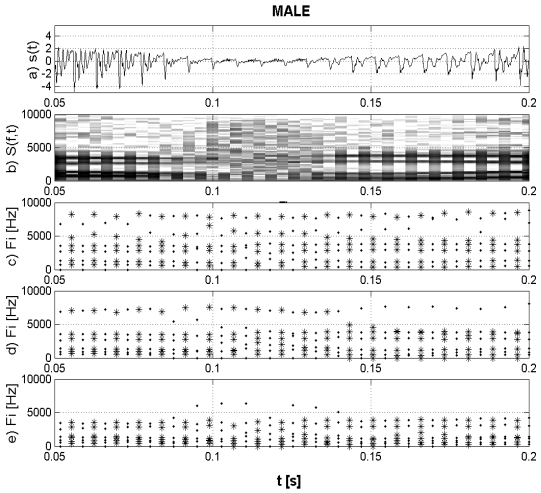

Figure 1. *a) Phonetic register /aveller/ by man. b) Spectrogram. c) WCPCZ ,λ=0.25. d) WCPCZ,λ=0.45. e) WCPCZ, λ=0.65.*

We can observe that the formant track the spectrogram accurately (fig. 2.b), with better resolution in lower frequencies than the non warped method ECPCZ (fig. 2.c.) In higher frequencies the performance is very acceptable if we bear in mind that the spectrogram does not provide perceptual information. We also can see, as its behaviour is more regular than the classical MCC (fig. 2.e). The formants of the open and closed phases follow the spectrogram with their own differences and they are able to define the transition clearly.

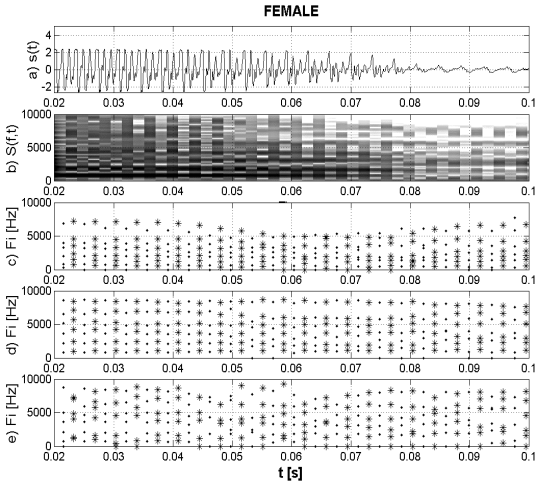

Figure 2. *a) Phonetic register /an/ by woman. b) Spectrogram. c) WCPCZ estimation, λ=0.45. d) ECPCZ estimation. e) MCC estimation.*

In figure 3.a we show the voice segment /ndmei/ uttered by a man which contains a consonant-consonant, consonant-vocal and vocal-vocal transitions. In the figure 3.c we observe the estimations of the WCPCZ method. We also can see an efficient tracking of the formants in relation to the spectrogram (fig. 3.b) and a good consistency for both the transition and the diphthong. The warped method has a better performance than the non warped estimations, carried out by ECPCZ (fig. 3.d) and MCC (fig. 3e).

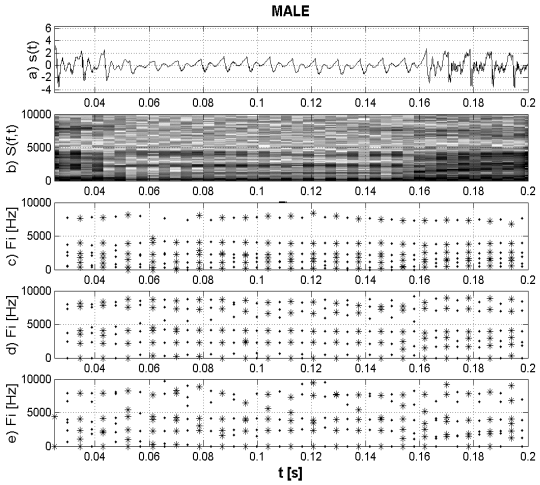

Figure 3. *a) Phonetic register /ndmei/ by man. b) Spectrogram. c) WCPCZ estimation, λ=0.45. d) ECPCZ estimation. e) MCC estimation.*

Finally, we use the methods in a moderate noisy environment, with 20 dB of noise a signal relation and a maximum of 3 samples of random error in the GCI's. The figure 4 shows the first example into the Gaussian noise, the voice segment /aveller/. The noisy spectrogram (fig. 4.b) can be compared with the clean spectrogram of the figure 1.b. We can realise the estimations of the proposed method WCPCZ over noisy signal (fig. 4.c) are well done. They are similar to the estimations achieved over the clean signal (fig. 1.d) and better than those obtained by non warped methods (fig 4.d y 4.e), specially in lower frequencies. The better resolution of the warped method in that band is the reason of its robustness, because of the high signal level and the accurate of the GCI not have to be so strict.

5. Conclusions

We have approached the problem of common voice parameters in several consecutive periods from the auditory perception point of view to track the speech natural phonetics variations. We adopted a formulation that is a good frame to define different approaches of coefficients estimates associated to polo and zero structures. This is good for both open and closed phases. The experiments have proved that the integration of the several periods with perceptual information improves the consistency and the robustness against moderate noise and provides greater reliability to both men and women, and also in phonetic transitions of a different nature.

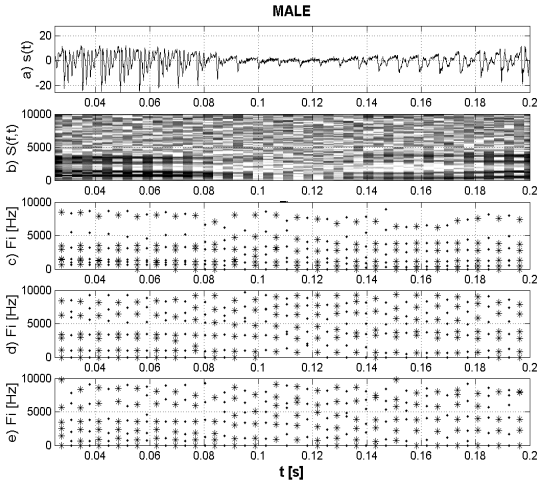

Figure 4. *a) Noisy phonetic register /aveller/ by man. b) Spectrogram. c) WCPCZ estimation, λ=0.45. d) ECPCZ estimation. e) MCC estimation.*

6. Acknowledgements

The authors would like to thank the Department of Communications and Neuroscience, Keele University, U.K., for supplying the speech database.

This work has been partially supported by the R&D National Projects TEC2004-09615-C03, TEC2005-08377-C03 and TEC2005-07010-C02.

7. References

[1] B. Yegnanarayana, R. N. Veldhuis. "Extraction of Vocal-Tract System Characteristics from Speech Signals*". IEEE Trans. on SAP*. July 98. Vol. 6. No 4, pp 313-327.

[2] H. Matsumoto, M. Moroto. "Evaluation of MEL-LPC Cepstrum in a Large Vocabulary Continuous Speech Recognition". *Proceedings of the IEEE-ICASSP*. Vol. 1, 7-11 May 2001, pp 117-120.

[3] A. Härmä, U.K. Laine, M Karjalainen. "An Experimental Audio Codec Based on Warped Linear Prediction of Complex Valued Signals". *Proceedings of the IEEE-ICASSP*. Vol. 1 , 21-24 April 1997, pp 323-326.

[4] P.J. Quintana-Morales, J.L. Navarro-Mesa. "An Approach to Common Acoustical Pole and Zero Modeling of Consecutive Periods of Voiced Speech". *Proceedings of the EuroSpeech*. 1-4 September 2003, pp 2433-36.

[5] P.J. Quintana-Morales, J.L. Navarro-Mesa. "Frequency Warped ARMA Modeling of the Closed and Open Phase of Voiced Speech". *Proceedings of International Conference on Spoken Language Processing–INTERSPEECH*, 4-8 October 2004.