An Approach to Common Acoustical Pole and Zero Modeling of Consecutive Periods of Voiced Speech

Pedro Quintana-Morales & Juan L. Navarro-Mesa
Departamento de Señales y Comunicaciones – Universidad de Las Palmas de Gran Canaria
Campus Universitario de Tafira – 35107 Las Palmas de G. C. – Spain
{pedroq,navarro}@dsc.ulpgc.es

Abstract

In this paper the open and closed phases within a speech period are separately modeled as acoustical pole-zero filters. We approach the estimation of the coefficients associated to the poles and zeros by minimizing a cost function based on the reconstruction error. The cost function leads to a matrix formulation of the error for two time intervals where the error must be defined. This defines a framework that facilitates to model the phases associated to consecutive periods. We give a matrix formulation of the estimation process that let us to attain two main objectives. Firstly, estimate the common-pole structure of several consecutive periods and their particular zero structure. And secondly, estimate their common-pole-zero structure. The experiments are carried out over a speech database of five men and five women. The experiments are done in terms of the reconstruction error and its dependence on the period length and the order of the analysis.

1. Introduction

The main objective of this paper is to propose a method to track natural variations of the vocal-tract common acoustical pole and zero structure in consecutive periods. This is a challenging problem which is crucial for applications in speech recognition, synthesis, segmentation, etc [1].

The pole and zero structure of the vocal tract varies in time in two distinct ways. First, The movement of the articulators change the shape of the tract. And second, the vocal folds oscillate between an open and a closed phase within each period even though the articulators themselves do not move. Except for some phonetic transitions (e.g., vowel-consonant), from period to period the natural variations are very slow.

Our study covers both variations. That is, we assume that within a small number of periods the movement of the articulators and the phase characteristics are constant or almost constant. We make a distinction between the instants of pre- (open phase) and postexcitation (closed phase) [2]. To face this study we will do a pitch-synchronous analysis instead of the typical block-processing one because the later smears the information of interest. The instants of main excitation are the instants of glottal closure (ICG).

An analysis in this way poses the problem of consistency in the estimations. This is because periods, and their corresponding phases, may be very too short. To overcome this problem, while achieving reliability in the estimates, we adopt the solution of using samples from consecutive periods (e.g. [2]). Our aim will be to integrate the information from these periods by looking for minimum reconstruction errors.

In section 2 we introduce a general formulation of the estimation error in matrix notation where the speech and excitation signals, and the parameters are explicitly separated. Also we apply a cost function to be minimized. In section 3 we propose to apply the previous formulation to estimate the common-pole and the particular zero structure. Going a step further in the assumption of slow natural variations, in section 4 we estimate the common-pole-zero structure to all periods. Then the cost function is generalized and we give a closed expression for the parameters that minimize it in the least-squares (LS) sense.

2. Pole and Zero Modeling of a Single Period

Let *y(n,k)* be the signal associated to a given phase (open or closed) within the *n-th* period *s(n,k)*. The expression for the pole-zero and 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, *{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 the Z domain, the poles and zeros of the acoustical filters are represented by the roots of the polynomials $A(z) = 1 + \sum_{i=1}^{p} a_i z^{-i}$ and $B(z) = \sum_{i=0}^{q} b_i z^{-i}$.

Let's now define the reconstruction error signal as:

$$e(n,k) = y(n,k) + \sum_{i=1}^{p} a_i^n y(n,k-i) - \sum_{i=0}^{q} b_i^n u(n,k-i) \qquad (2)$$

Using a matrix notation (e.g., [3]), and assuming that *u(n,k)* is known or appropriately estimated, equation (2) is:

$$\underline{e_n} = \underline{y_n} - \left[\underline{\underline{Y_n}} \quad \underline{\underline{U_n}}\right]\underline{h_n} = \underline{y_n} - \underline{\underline{H_n}}\,\underline{h_n} \qquad (3)$$

$$\underline{e_n} = \left[e(n,0), \cdots, e(n,N_n-1)\right]^T \ , N_n x 1$$

$$\underline{y_n} = \left[y(n,0), \cdots, y(n,N_n-1)\right]^T \ , N_n x 1$$

$$\underline{h_n} = \left[a_1^n \ \cdots \ a_p^n \ b_0^n \ \cdots \ b_q^n\right]^T \ , (p+(q+1)) x 1$$

$$\underline{\underline{Y}}_n = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ y(n,0) & 0 & \cdots & 0 \\ y(n,1) & y(n,0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ y(n,p-1) & y(n,p-2) & \cdots & y(n,0) \\ \vdots & \vdots & \ddots & \vdots \\ y(n,N_n-2) & y(n,N_n-3) & \cdots & y(n,N_n-p-1) \end{bmatrix}, N_n x p$$

$$\underline{\underline{U}}_n = \begin{bmatrix} u(n,0) & 0 & \cdots & 0 \\ u(n,1) & u(n,0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u(n,q) & u(n,q-1) & \cdots & u(n,0) \\ \vdots & \vdots & \ddots & \vdots \\ u(n,N_n-1) & u(n,N_n-2) & \cdots & u(n,N_n-q-1) \end{bmatrix}, N_n x(q+1)$$

This equation leads to a classical Least-Squares method of estimation. Paying attention to the equation in (3) one can notice that $\underline{e}_n$ is defined over an interval *{k=0,,...,N_n-1}*. Making a few manipulations it is easy to show that the interval can be defined over a larger range because the signal $\underline{\underline{Y}}_n$ and the excitation $\underline{\underline{U}}_n$ matrixes can be extended due to the infinite impulse response inherent in the ARMA model. Then, we can think about taking advantage from the usage of the whole interval over which the error is different from zero. With this idea in mind, we propose to redefine matrixes as follows:

$$\underline{e}_n = \left[e(n,0),\cdots,e(n,N_n+p-1)\right]^T, (N_n+p)x1 \qquad (4)$$
$$\underline{y}_n = \left[y(n,0),\cdots,y(n,N_n-1),0,\cdots,0\right]^T, (N_n+p)x1$$
$$\underline{h}_n = \left[a_1^n \cdots a_p^n \; b_0^n \cdots b_q^n\right]^T, (p+(q+1))x1$$

$$\underline{\underline{Y}}_n = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ y(n,0) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ y(n,N_n-2) & y(n,N_n-3) & \cdots & y(n,N_n-p-1) \\ y(n,N_n-1) & y(n,N_n-2) & \cdots & y(n,N_n-p) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y(n,N_n-1) \end{bmatrix}, (N_n+p)xp$$

$$\underline{\underline{U}}_n = \begin{bmatrix} u(n,0) & 0 & \cdots & 0 \\ u(n,1) & u(n,0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u(n,N_n-1) & u(n,N_n-2) & \cdots & u(n,N_n-q-1) \\ 0 & u(n,N_n-1) & \cdots & u(n,N_n-q) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix}, (N_n+p)x(q+1)$$

Equation (4) is valid when *(q+1)<p*. In the case that *(q+1)≥p,* this equation can be easily rearranged and takes a similar form. Equation (4) leads to what we call Extended Error Least-Squares .

Irrespective of the formulation of the matrixes $\underline{\underline{Y}}_n$ and $\underline{\underline{U}}_n$ in (3) and (4) we define the following cost function:

$$C_1(n) = \sum_{k=0}^{L} e^2(n,k) \qquad (5)$$

where the upper limits of summation *L* are *(N_n-1)* and *(N_n+p-1)* for equations (3) and (4), respectively. It is easy to see from the matrix counterpart of (5) [3] that the coefficients that minimize *C_1(n)* using the LS method is:

$$\underline{h}_n = (\underline{\underline{H}}_n^T \underline{\underline{H}}_n)^{-1} \underline{\underline{H}}_n^T \underline{y}_n \qquad (6)$$

3. Common-pole and Zero Modeling from Consecutive 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. Then, we can assume that irrespective of the phase, for some *M* consecutive periods of speech the zero structure slightly varies from period to period, not in a significant way. We also assume that the pole (formant) structure is constant. In this case it is possible to redefine equations (3) and (4) in order to make an estimation of the coefficients. Now, the error equation takes the form:

$$\underline{e} = \underline{y} - \underline{\underline{H}}\,\underline{h}_M \qquad (7)$$

$$\underline{\underline{H}} = \begin{bmatrix} \underline{\underline{Y}}_0 & \underline{\underline{U}}_0 & \underline{\underline{0}} & \cdots & \underline{\underline{0}} \\ \underline{\underline{Y}}_1 & \underline{\underline{0}} & \underline{\underline{U}}_1 & \cdots & \underline{\underline{0}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \underline{\underline{Y}}_{M-1} & \underline{\underline{0}} & \underline{\underline{0}} & \cdots & \underline{\underline{U}}_{M-1} \end{bmatrix}$$

$$\underline{e} = \left[\underline{e}_0,\cdots,\underline{e}_{M-1}\right]^T, (MxN_n)x1$$
$$\underline{y} = \left[\underline{y}_0,\cdots,\underline{y}_{M-1}\right]^T, (MxN_n)x1$$
$$\underline{h}_M = \left[\underline{a},\underline{b}^0,\cdots,\underline{b}^{M-1}\right]^T, (p+Mx(q+1))x1$$
$$\underline{a} = \left[a_1 \cdots a_p\right]^T \qquad \underline{b}^j = \left[b_0^j \cdots b_q^j\right]^T$$

where $\underline{\underline{Y}}_j$ and $\underline{\underline{U}}_j$, *{j=0,...,M-1}*, are the signal and excitation matrixes corresponding to a given phase *y(n+j,k)* of the *s(n+j,k)* period. The signal $\underline{y}_j$ and error $\underline{e}_j$ vectors are similar to the ones in (3). The firsts, *{a_i}*, correspond to the common pole structure and the rest, *{b_i^j}*, correspond to their particular zero structure.

This equation leads to what we call the (Extended) Common-Pole and Particular Zero over *M(≥1)* periods (E)CPPZM.

4. Common-pole-zero Modeling from Consecutive Periods

In this section we assume that both pole and zero structures are constant during the M consecutive periods. Equation (7) needs some modifications to take into account the fact that now the coefficients $\{b_i^j\}$ are common to all periods under analysis.

$$\underline{\underline{H}} = \begin{bmatrix} \underline{\underline{Y_0}} & \underline{\underline{U_0}} \\ \underline{\underline{Y_1}} & \underline{\underline{U_1}} \\ \vdots & \vdots \\ \underline{\underline{Y_{M-1}}} & \underline{\underline{U_{M-1}}} \end{bmatrix}, (MxN_n)x(p+(q+1)) \qquad (8)$$

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

This equation leads to what we call the (Extended) Common-Pole Common-Zero over $M(\geq 1)$ periods (E)CPCZM.

Let's face the estimation of $\underline{h_M}$. Our error equation in (7) is similar to that in [4] where the authors estimate the common-acoustial-pole and zeros of several head-related transfer functions. Despite the differences in application, both situations share a similar matrix formulation that leads to a solution using the least-squares method. The cost function is now defined as the square sum of the reconstruction error for time index k of the signal of a given phase within M consecutive periods starting from the *n-th* one.

$$C_M(n) = \sum_{j=0}^{M-1} \sum_{k=0}^{L} e^2(n+j,k) \qquad (9)$$

The coefficients $\underline{h_M}$ that minimize $C_M(n)$ in (7) and (8) using the least-squares method can be represented now in vector form as:

$$\underline{h_M} = (\underline{\underline{H^T H}})^{-1} \underline{\underline{H^T}} \underline{y} \qquad (10)$$

The authors in [4] probe this solution to be adequate for the estimation of the optimum orders (p,q). Instead, we adopt this solution to study the validity of equations (6) and (10) to track the natural variations in the vocal tract.

Before proceeding to the experiments, some considerations must be done; i), the interval in which $\underline{e_n}$ is defined can be done in two ways as given in (3) and (4), ii) equations (7) to (10) are equivalent to equations (3) to (6) for $M=1$, and iii) for the sake of simplicity we have implicitly considered that N_n in (7) and (8) are constant over M.

5. Experiments and Results

We have used a speech database with their corresponding laryngogram of 5 men and 5 women (each one is about 40 seconds long). The sampling frequency is 20 KHz. Prior to the experiments we have used the laryngogram signals to mark the correct IGC (29292 were obtained), voiced/unvoiced intervals and pitch. From the ICG the open and the closed phases are extracted. In each period, the closed and opened phases represent the 40% and 48% of each period length, respectively. As suggested in [2] the open phase ends an arbitrary (12% is our compromise value) instant before the excitation.

For figures 1 and 2, the (p,q) couples we have studied vary within the following ranges; $(1 \leq p \leq 20)$, $(1 \leq q \leq 19)$. These orders seem a good choice for the maximum frequency in th signals (10Khz). Given a $(p,q,)$ order, the error is computed as the average from the open and closed phases from all periods. In figure 1 we show the dependence on the (p,q) model order of reconstruction error over one period when using the classical and the extended formulation in (3) to (6). We can see that the extended error formulation (Fig. 1a) produce smaller error for any given (p,q) than the classical one (Fig. 1b). This improvement in reconstruction error is at the expence of a higher dimensionality of the matrixes in the extended method. However, the increase in dimensionality is not computationally prohibitive.

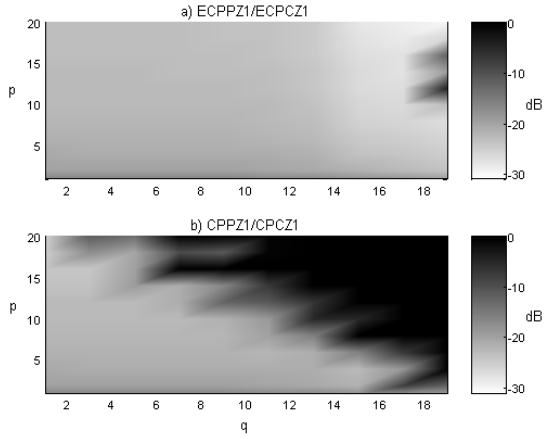

Figure 1. Reconstruction error (dB) for the extended (a) and classical (b) methods in (6) for 1 period

In figure 2 we show the dependence on the (p,q) model order of reconstruction error over $M=3$ periods when using the classical and the extended formulation in (7) to (10). It can be appreciated that in general the extended error formulation achieves smaller errors than the classical one. This is specially true for the Common-Pole and Particular Zero (Fig. 2a and 2b). Also, compared with the results in figure 1 the errors are smaller too. This reinforces the idea that the formulation we propose for integrating several periods is appropriate to estimate the common parameters. We have not specifically studied consonat-vowel transitions. In these particular situations the assumption of slow variations in

the vocal tract fails and the errors are expected to be higher than in other transitions.

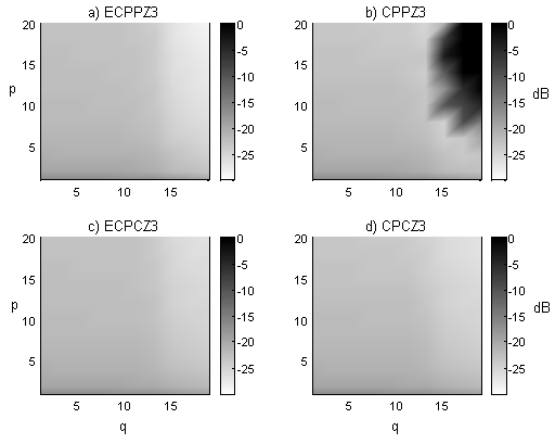

Figure 2. Reconstruction error (dB) for extended (a) and classical (b) CPPZ, and extended (c) and classical (d) CPCZ, methods in (10) for *M*=3 periods

In figure 3 we show the dependence on the period length of the reconstruction errors over one period (closed and open phases separatedly) when using the classical and the extended formulation of the error in (6) for *p=18, q=17*. This *(p,q)* orders are the ones that pose the worst problem in terms of consistensy for short periods. As we can see, the main advantage of the extended method over the classical is that for small period lengths (≤5 ms) the reconstruction error is clearly smaller. For higher periods the error is on the same order of magnitude.

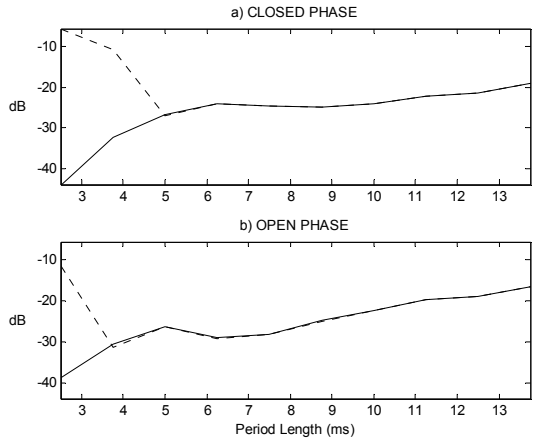

Figure 3. Reconstruction error (dB) for the extended (-) and the classical LS (- -) methods in (6) for 1 period

In figure 4 we show the dependence on the period length of the reconstruction errors over *M=3* periods (closed and open phases separatedly) applying the classical and the extended error formulation in (10) when we apply the (E)CPPZM and the (E)CPCZM, for *p=18, q=17*. Again, the overall error for the classical error formulation is higher than the one for the extended when the period is small. On the other hand, the error for the (E)CPPZM is slightly lower than that for the (E)CPCZM.

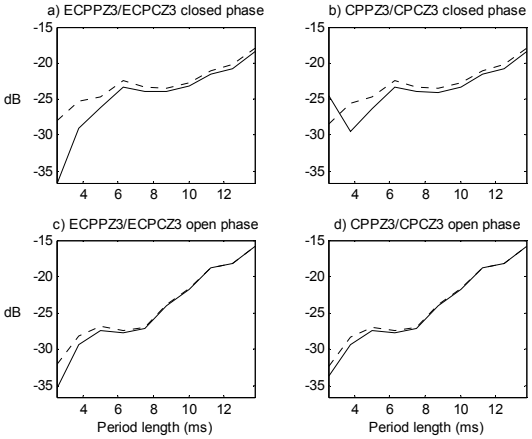

Figure 4. Reconstruction error (dB) for closed a)- b), and open c)-d) phases using the extended ECPPZ3 (-)/ ECPCZ3 (- -) and the classical CPPZ3 (-)/CPCZ3(- -)

6. Conclusions

We have addressed the problem of estimating the common parameters of speech in consecutive periods. The matrix formulation we adopt is a good framework to define several approaches to the estimation of the coefficients associated to the pole and zero structure. It is valid for both open and closed phases. The experiments show that the extended formulation in (4) improves the one in (3) for *M=1* or higher in terms of reconstruction error for any given *(p,q)*. Also, the dependence on the period length is better with the extended error formulation. In future work we will study more in detail aspects like consistency and reliability.

7. References

[1] L. Rabiner, R. Schafer. "Digital Processing of Speech Signals". Englewood Clifs, NJ. Prentice-Hall, 1978.

[2] B. Yegnanarayana, R. N. Veldhuis. "Extraction of Vocal-ract System characteristics from Speech Signals". IEEE Transactions on SAP. July 98. Vol. 6. No. 4, pp 313-327.

[3] S. M Kay."Modern Spectral Estimation: Theory and Application".Prentice-Hall Signal Proc. S., 1988.

[4] Y. Haneda, S. Makino, Y. Kaneda, N. Kitawaki. "Common-Acoustical-Pole and Zero Modelling of Head-Related Transfer Functions". IEEE Trans. on SAP. March 1999. Vol. 7. No. 2.