Modified Realizable Frequency Warped ARMA Modeling and its Application in Synthesis Structures for Voiced Speech

Juan L. Navarro-Mesa & Pedro J. Quintana-Morales
Departamento de Señales y Comunicaciones
Universidad de Las Palmas de Gran Canaria – Spain
{jnavarro, pquintana}@dsc.ulpgc.es

Abstract

Synthesis filters are of major concern in many speech applications. In this paper these filters are designed so as to track the natural variations, parameterize the perceptually relevant aspects of the pole and zero characteristics and their dynamics, and achieve reliable estimates of the filter coefficients. In this paper this is attained integrating three basic points of view in a general framework. First, frequency warping incorporates the perceptually relevant characteristics of human hearing. Second, direct implementations of efficient synthesis structures of these filters are realized with modified versions that overcome the problems associated to the incorporation of first-order all-pass sections for frequency warping. And third, reliable coefficients associated to the synthesis structures of several consecutive periods are directly estimated in the analysis. We show that the synthesis structures we apply and their coefficients offer good signal to reconstruction error ratios and error distribution in frequency.

1. Introduction

Understanding and modeling the natural variations in the characteristics of the vocal-tract system may be crucial in speech coding, synthesis and recognition (e.g., [1,2]). These variations can be analysed from the pole and zero characteristics of speech. In voiced speech these characteristics vary in two distinct ways. First, the movement of the articulators change the shape of the vocal tract. And second, the vocal folds oscillate between open and closed phases within each period even though the articulators themselves do not move. Except for some phonetic transitions from period to period the natural variations are slow. In our study we perform a pitch-synchronous analysis where periods are delimited by the instants of glottal closure. We distinction between pre (open) and postexcitation (closed) phases.

Regardless phase under analysis, we will model the signal as an ARMA process whose expression is

\[ y(n,k) = \sum_{i=1}^{p} \alpha_i^y u(n,k-i) + \sum_{i=1}^{q} \beta_i^y d(n,k) \]  

where \( u(n,k) \) is the excitation signal within the n-th period, \( \{k=0,\ldots,N_n-1\} \), \( N_n \) is the phase length and \( \{\alpha_i^y, i=1,\ldots,p\} \) and \( \{\beta_i^y, i=0,\ldots,q\} \) are the AR and MA coefficients of orders (p,q), respectively. In a previous work [3] we have developed a method for estimating the common acoustical pole and zero structure to several consecutive periods. This method has the ability to offer good reconstruction error signals.

Incorporating an auditory perception point of view in the analysis stage has revealed important in coding [4] and recognition [5] as well. This lets us to address the problem of tracking the natural variations of consecutive periods of voiced speech with a formulation that is a compromise between two aims [6]. These are, error minimisation in the temporal domain and in the spectral domain where it were more disturbing (e.g., at low frequencies). The expression in the warped frequency scale for the ARMA process is [7]

\[ y(n,k) = \sum_{i=1}^{p} \alpha_i^y d(y(n,k)) + \sum_{i=1}^{q} \beta_i^y d(u(n,k)) \]  

where \( d(n) \) is the impulse response of a first-order all-pass dispersive section that substitutes the unit delay element \( z^{-1} \) in (1) and \( d(y(n,k)) \) is the all-pass section output in the i-stage of a cascade with input signal \( y(n,k) \). The dispersive section depends on a factor \( \lambda \) that controls the frequency warping (e.g., Bark, Mel) [8].

Practical implementations of the frequency-warped filters pose realization problems since the first-order all-pass sections contain delay-free loops. This means that frequency warping as applied in [4,5,6] etc. are rather useful for analysis than for synthesis. Then, the question is whether it is possible to have realizable structures. The modified structures proposed in the literature (e.g.,[9,10,11]) bring the opportunity of integrating filter’s realizability and coefficient estimation in a joint mathematical formulation.

2. Modified Realizable Structures

A general form for the transfer function of warped IIR filters is

\[ H(z) = \frac{\sum_{i=1}^{p} \beta_i^D(z)}{1 + \sum_{i=1}^{p} \alpha_i^D(z)} \]  

where \( (p,q) \) are the AR and MA orders, respectively. This is what we call the ideal warped structure (figure 1) where first order all-pass sections \( D(z) \) substitute unit delay elements. A realization problem appears since delay sections contain delay-free components as seen in

\[ D(z) = \frac{1 - \lambda^2 z^{-1}}{1 - \lambda z^{-1}} \]  

Figure 1: Ideal non-realizable structure (InRS)
This problem can be overcome with a structure of the form in figure 2 where in the IIR part the first section is now a lowpass filter \( \{g(n), G(z)\} \) and the other allpass sections are \( D(z) \). This is a robust realizable structure where the AR part was proposed by Steiglitz using IIR structures direct form I. The FIR part remains unchanged while the IIR part has a modified structure in which delay-free loops are eliminated. The new IIR (AR) coefficients \( \eta_i \) can be estimated from the originals \( \alpha_i \) with recursive formulas [10].

![Figure 2: Realizable Steiglitz structure (RSS)](image)

A different robust realizable structure [11] is presented in figure 3 where feedback coefficients \( \alpha_i \) are mapped to coefficients \( \sigma_i \) which feed back from the outputs of the unit delays of all-pass sections in order to avoid lag-free loops. As in the Steiglitz the feedforward (FIR) part can be implemented directly and the mapped coefficients can be estimated from recursive formulas [11].

The added complexity of the warped structures, when compared with the ordinary FIR and IIR, makes interesting to propose new ones where the complexity is lower. In figure 4 we present a slight modification of the structure in figure 2.

![Figure 3: Realizable mapped structure (RMS)](image)

![Figure 4: Modified realizable Steiglitz structure (MSS)](image)

This new structure is a compromise between the ideal non-realizable in figure 1 and the realizable mapped in figure 3. Notice that in the IIR part the first allpass section is the unit delay element \( g(n) = \delta(n-1), G(z) = z^{-1} \) and the other allpass sections are \( D(z) \). The objective is to improve efficiency while approaching the idealization expressed in figure 1 in a realizable structure.

Up to now we have not considered the estimation of \( \eta_i \) or \( \sigma_i \). This is because we are really interested in estimating these coefficients directly from the signal without a transformation from the ideal ones. In the next section we will study this goal.

### 3. Least-Squares Coefficient Estimation for Several Consecutive Periods

Before deriving the estimation formulas let’s recall the equation in (2) and particularize for the different structures. Then, the expressions for the synthesis signals are:

- **RSS**
  \[
  y^{RSS}(n,k) = \frac{1}{\alpha_0} \left[ -\sum_{i=1}^{p} \eta_i g(y(n,k)) - \sum_{i=2}^{\infty} \sum_{l=2}^{\infty} \beta_{i,l} d_i (u(n,k)) \right]
  \]

- **MSS**
  \[
  y^{MSS}(n,k) = \frac{1}{\sigma_0} \left[ -\sum_{i=1}^{q} \sigma_i y(n,k-1) - \sum_{i=2}^{\infty} \sum_{l=2}^{\infty} \beta_{i,l} d_i (u(n,k)) \right]
  \]

- **RMS**
  \[
  y^{RMS}(n,k) = \frac{1}{\sigma_0} \left[ -\sum_{i=1}^{q} \sigma_i y(n,k-1) - \sum_{i=1}^{\infty} \sum_{l=1}^{\infty} \beta_{i,l} d_i (u(n,k)) \right]
  \]

This is a special case because it is necessary to introduce intermediate variables \( s_i(n) \) and \( r(n) \). These are:

\[
\begin{align*}
  s_i(n) &= s_{i-1}(n-1) + \lambda(s_i(n-1) - s_{i-1}(n)) \\
  r_i(n) &= r_{i-1}(n-1) + \lambda(r_i(n-1) - r_{i-1}(n)) \\
  z_j(n) &= z_{j-1}(n-1) + \lambda(z_j(n-1) - z_{j-1}(n)) \\
  z_i(n) &= z(n)
\end{align*}
\]

and the expression for the synthesis signal is:

\[
  y^{RMS}(n,k) = \frac{1}{\sigma_0} \left[ \sum_{i=1}^{q} \sigma_i \sum_{l=1}^{\infty} \beta_{i,l} r_i(n,k) + \sum_{i=1}^{\infty} \sum_{l=1}^{\infty} \beta_{i,l} d_i (u(n,k)) \right]
\]

\[
  y^{RSS}(n,k) = \frac{1}{\alpha_0} \left[ \sum_{i=1}^{p} \eta_i \sum_{l=2}^{\infty} \beta_{i,l} d_i (u(n,k)) \right]
\]

\[
  y^{MSS}(n,k) = \frac{1}{\sigma_0} \left[ \sum_{i=1}^{q} \sigma_i \sum_{l=2}^{\infty} \beta_{i,l} d_i (u(n,k)) \right]
\]
Let's now define the reconstruction error signal as
\[ e_n = y_n - \frac{Y}{U_f} h_n = y_n - H_{\mu} h_n \]  
(9)

\[ Y = \begin{bmatrix} g(y(n,0)) & d_f(y(n,0)) & \cdots & d_p(y(n,0)) \\ g(y(n,1)) & d_f(y(n,1)) & \cdots & d_p(y(n,1)) \\ \vdots & \vdots & \ddots & \vdots \\ g(y(n,N_n-1)) & d_f(y(n,N_n-1)) & \cdots & d_p(y(n,N_n-1)) \\ g(y(n,N_n+N'-1)) & d_f(y(n,N_n+N'-1)) & \cdots & d_p(y(n,N_n+N'-1)) \end{bmatrix} \]

where \( h_n, y_n \) and \( e_n \) are the coefficients, the signal and the error vectors, respectively. And \( Y \) and \( U_f \) are the signal and excitation matrices, respectively. For the sake of simplicity we do not show the expression of \( Y \) and \( U_f \) for RMS. These are similar to (9) but applying \( z_i \) and \( r_i \) instead of \( d_i(u(n)) \) and \( d_i(y(n)) \), respectively. In the case that \((q+1)p\), the equations can be easily rearranged and takes similar forms. We have introduced an extension parameter \( N' \) that controls the time intervals in samples where the error is defined.

Now we are ready for extending the formulation in (9) to the case in which the estimation of the common poles and zeros of \( M \) consecutive periods is of concern. Then we can write

\[ e = y - H_{\mu} h_{\mu} \]  
(10)

\[ H = \begin{bmatrix} Y_0 & U_0 & 0 & \cdots & 0 \\ Y_1 & 0 & U_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{M-1} & 0 & 0 & \cdots & U_{M-1} \end{bmatrix} \]

\[ y = \begin{bmatrix} y_1^T \\ \vdots \\ y_{N+1}^T \end{bmatrix} \quad h_{\mu} = \begin{bmatrix} \phi_1^T \\ \vdots \\ \phi_{N+1}^T \end{bmatrix} \quad \sigma = \begin{bmatrix} \sigma_1 & \cdots & \sigma_{N+1} \end{bmatrix} \]

\[ h_{\mu} = (H^T H)^{-1} H^T y \]  
(11)

3. 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 \( f_{\text{ftp}} \) 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 intervals and pitch. From the ICG the open and the closed 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. In all experiments \( M=3, \lambda=0.75 \) and \( N'=20 \). Also, the \((q,p)\) order is set to \((15,16)\). This seems a good choice if we consider that in 10 KHz there are about 8 formants.

In the first experiment we show an example of error distribution in frequency. The figure 5 corresponds to a segment of three long periods (about 13 milliseconds).

The spectrum is in continuous line and the reconstruction errors are in dotted lines. It is not possible to distinguish between any of the errors from the different structures. The maximum effort for minimizing the error is concentrated in frequencies lower than the turning point frequency \( f_p \) defines the frequency for which warping does not affect the frequency resolution, i.e., where the group delay is one. In this experiment \( f_p = 2305 \) Hz. For higher frequencies the effort in minimizing the error decreases until becoming irrelevant about 4 KHz [7].
In the next experiment we show (figure 6) the signal to reconstruction error relation (SRR) in dB evaluated over all periods. The upper (dashed) curve corresponds to the ideal structure InRS and the lower [3] corresponds to the non-warped structure as in (1). The middle curves correspond to the RSS, MSS and RMS structures. As we can see the best SRR corresponds to the ideal case. However, this is not realistic since InRS is only valid for analysis. When designed for synthesis it has problems with delay-free loops as stated in the previous sections.

## 4. Concluding Remarks

Realizable warped synthesis filters have not received enough attention for direct estimation of the coefficients. In this paper we have introduced a general framework which is appropriate for tracking the natural variations of voiced speech while preserving the main perceptual characteristics. From an implementation point of view we have developed a matrix formulation for estimating the filter’s coefficients in the least-squares sense. With this formulation we have approached the problem of tracking the pole and zero characteristics of voiced speech. It is interesting to notice that none structure offers significant performance in terms of reconstruction error. The main differences come from computation time.

## 5. References