Proc. IEEE Workshop Speech Coding (Pocono Manor, PA), pp. 1-2, Sept. 1997

A NEW LPC ERROR CRITERION FOR IMPROVED PITCH TRACKING

Mohammad R. Zad-Issa and Peter Kabal

Electrical Engineering, McGill University, Montreal, Quebec, Canada H3A 2A7

ABSTRACT

In Linear Predictive coders the output of the LP analysis filter is used to represent the glottal excitation signal. For high pitched voices during nasal sounds or nasalized vowels, the speech signal takes on a sinusoidal shape. The corresponding residual signal has a very low energy and the pitch pulses are weak or absent, resulting in poor pitch tracking. These segments of speech are also characterized by large frame-to-frame variations of the LP coefficients. In this paper we propose a composite formant prediction error criterion leading to a clear track of residual pulses even for for the sinusoid-like speech, while enhancing the smoothness of the filter parameter evolution.

1. INTRODUCTION

For voiced speech, Code Excited Linear Predictive Coders (CELP) model the excitation signal by selecting the "best" waveform from an adaptive codebook containing past pitch pulses. To accomplish the same task many of the more recent proposed coders such as Waveform Interpolation (WI) [1] model the pitch pulse shape. All of these coders use linear prediction to model the vocal tract excitation signal. Since appropriate modeling of the excitation signal directly affects the quality of the output speech, the success of the pulse coding stage depends on the closeness of the excitation model to the true glottal signal.

High pitched speech during nasals and nasalized sounds often takes on a sinusoidal form. In the next section, we will show that in addition to having large frame-to-frame fluctuations of the LP parameters, these segments are characterized by having a very low energy residual in which the pitch pulses are nearly absent. This signal is no longer a good representation of the true excitation to the vocal tract. Analysis-by-Synthesis coders may not select the appropriate excitation segment, consequently, the coding efficiency and the speech quality degrades. Also, in multi-mode coders if the voicing decision is based on the LP residual then these segments of speech may be misidentified as unvoiced speech. In this work, we add a second term to the conventional LP error criteria to account for the smoothness in the evolution of LP parameters. The contribution of this second term to the overall error function is controlled by the numerical conditioning of the correlation matrix. This modification avoids the disappearance of pitch pulses from the residual signal.

This work was supported by the Fonds pour la Formation de Chercheurs et l'Aide à la Recherche (FCAR)

2. COMPOSITE FORMANT PREDICTION ERROR CRITERIA

Nasal sounds are characterized by a low first formant (near 250 Hz) which dominates the power spectrum. The anti-resonance due to the closed oral cavity results in a weak second formant. For these nasals or nasalized vowels, when the harmonics are widely spaced (i.e. high pitched speech), the concentration of energy in low frequencies and the presence of spectral zeros may leave only one or two dominant harmonics in the signal power spectrum. This explains the sinusoid-like form of the speech signal during these segments. For a pure sinusoidal signal it can easily be shown that the rank of the correlation matrix $\mathbf{R}$ is two, therefore a second order predictor suffices to produce a zero residual. This does not hold when the input signal is windowed prior to the computation of $\mathbf{R}$, as is the case for the autocorrelation method. However, as the ratio of the analysis window length to the sinusoid period increases, the numerical rank[1] of $\mathbf{R}$ rapidly approaches two. For sinusoidal speech segments, experiments indicate that the numerical rank of the correlation matrix is between three and five (Fig. 1). The LP equations are essentially overdetermined if the prediction order is larger than the rank. Therefore, the solution is not unique and large frame-to-frame variations in the LP parameters may occur for such sinusoidal speech segments.

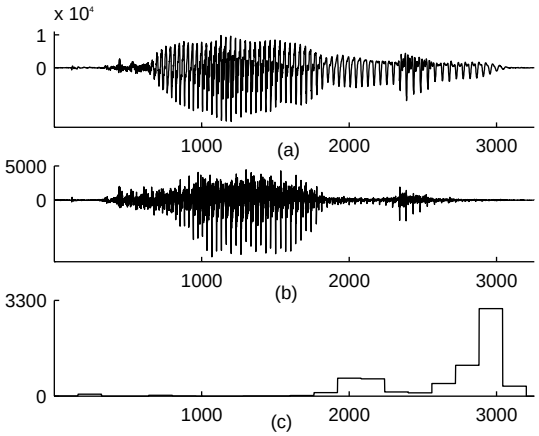

Figure 1: (a) Speech waveform. (b) Conventional LP residual. (c) λ_1/λ_4, λ_i is the i-th eigenvalue of $\mathbf{R}$.

In our method, we derive the formant prediction filter parameters by minimizing an error function containing two

[1] The numerical rank of a $n \times n$ matrix is k if $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \gg \lambda_{k+1} \geq \ldots \geq \lambda_n$. Where λ_i are the eigenvalues of the matrix.

terms. The first is the conventional LP error criterion, i.e. the energy of the output of the short term predictor, while the second term reflects the variation of LP coefficients with respect to those of the previous frame.

$$\begin{aligned}\mathbf{E} &= \mathbf{E}_{\mathrm{e}} + \mu \mathbf{E}_{\mathrm{a}} \\ &= (1 - 2\mathbf{r}'^{T}\mathbf{a} + \mathbf{a}^{T}\mathbf{R}'\mathbf{a}) + \mu(\mathbf{a} - \mathbf{a}_p)^{T}\mathbf{W}(\mathbf{a} - \mathbf{a}_p)\end{aligned} \quad (1)$$

Where $\mathbf{a}$ and $\mathbf{a}_p$ are the filter parameters for the current and the previous frame, $\mathbf{W}$ is a weighting matrix, $\mathbf{R}'$ and $\mathbf{r}'$ are the normalized autocorrelation matrix and vector, respectively. The scaler μ is the weighting factor for the additional error term. Solving for $\mathbf{a}$ leads to

$$(\mathbf{R}' + \mu\mathbf{W})\,\mathbf{a} = (\mathbf{r}' + \mu\mathbf{W}\mathbf{a}_p) \quad (2)$$

One choice of $\mathbf{W}$ is the normalized autocorrelation matrix associated with the previous frame, $\mathbf{R}'_p$. The solution $\mathbf{a}$ therefore minimizes the prediction error for averaged correlation values. Another choice for $\mathbf{W}$ is the identity matrix. In this case, the error becomes a function of the energy in the difference between the impulse response of the short term predictor filter of consecutive frames. Experiments show that by adjusting the weight μ, nearly identical results are obtained for $\mathbf{W}$ set to $\mathbf{I}$ or $\mathbf{R}'_p$. This procedure has the side benefit that the pitch pulses in the residual become clearly visible and more alike.

Appropriate selection of the weight μ assures a well conditioned system of equations. However if μ is too large, the loss in short term prediction gain[2] becomes excessive. This suggests that the weight μ should be determined on a frame-to-frame basis, where its value increases with the spread of the eigenvalues of $\mathbf{R}'$. We choose μ according to the following smooth switching function:

$$\mu = \frac{\rho}{2}(1 + \tanh(\frac{\xi - \alpha}{\beta})) \quad (3)$$

Where ρ is a scaling factor, ξ measures the conditioning of $\mathbf{R}'$. The constant α is the average value ξ for an ill-conditioned system. The values of α and β are determined experimentally based on ξ and $\mathbf{W}$.

3. SIMULATION RESULTS

High pitch female speech was sampled at 8 kHz. Linear prediction coefficients were calculated according to Eq. (2) for 20 ms frames, using a 30 ms Hamming analysis window for the autocorrelation method. To filter the input speech, these parameters were linearly interpolated four times per frame. Adjustment of the weight μ requires the knowledge of the distribution of the eigenvalues of the correlation matrix. To sidestep the computational load associated with eigendecomposition, we compared the use of the DFT and DCT [2] to approximately diagonalize the correlation matrix. The simulations indicated that nearly identical results are obtained by using the DCT and eigendecomposition. To evaluate the performance of this new error criterion we monitor the LP prediction gain, the similarity between successive pitch pulses as measured by the prediction gain of a three tap pitch predictor updated every 5 ms, and the average of the 1-norm of the LP parameter difference vector in the LSF (ω) domain.

[2]The ratio of the energy at the output to the energy at the input of the filter in dB.

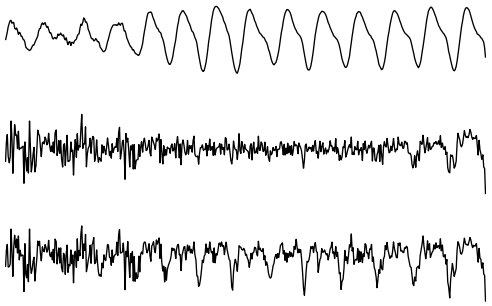

Figure 2: From top: Speech waveform, the conventional LP residual, the residual obtained using the new error criterion.

Prediction gain (dB)					
LP	Pitch	Overall	$\overline{\|\Delta\omega\|}_1$	α	$\mathbf{W}$
12.73	6.04	18.77	0.76	-	$\mathbf{0}$
12.70	6.04	18.74	0.65	400	$\mathbf{I}$
12.72	6.05	18.77	0.69	300	$\mathbf{R}'_p$

Table 1: Autocorrelation method, $\beta = 90$, $\rho = 1$, $\xi = \hat{\lambda}_1/\hat{\lambda}_4$, where $\hat{\lambda}_i$ is the DCT approximation to λ_i.

Figure 2 illustrates a sinusoid-like input signal. While the conventional residual signal contains very weak pitch pulses, the new residual displays clear pitch pulses. Additional testing with a pitch tracker shows it is able to follow the pitch track only in the second case.

Table 1 displays the results of the LP analysis for 7 s of high pitched speech, embedding several sinusoidal segments. As expected, the use of a smooth switching function to adjust μ guarantees that the overall prediction gain (sum of the LP and pitch prediction gains) remains almost unchanged. Also, the inter-frame variation in the LP coefficients for the new method has been reduced (smaller $\overline{\|\Delta\omega\|}_1$). Similar results are obtained when the covariance method is used for LP analysis.

4. CONCLUSION

In this paper we have presented a composite error measure to obtain the LP filter coefficients. This new criteria accounts for the prediction error and the evolution of the LP parameters. Using this approach, without a significant increase in the computational cost, we effectively increase the smoothness in the linear prediction parameters and prevent the disappearance of the pitch pulses for the sinusoid-like speech waveforms.

5. REFERENCES

[1] W. B. Kleijn and J. Haagen, "A speech coder based on decomposition of characteristic waveforms," *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Atlanta GA), pp. 508–511, 1995.

[2] M. Narasimha and A. M. Peterson, "On the computation of the discrete cosine transform," *IEEE Trans. Communications*, vol. 26, pp. 934–936, June 1978.