SMOOTHING THE EVOLUTION OF THE SPECTRAL PARAMETERS IN LINEAR PREDICTION OF SPEECH USING TARGET MATCHING

Mohammad R. Zad-Issa and Peter Kabal

Electrical Engineering Department, McGill University
3480 University Street, Montreal, Quebec, Canada H3A 2A7,
mohammad@tsp.ee.mcgill.ca
kabal@tsp.ee.mcgill.ca

ABSTRACT

Linear prediction (LP) coefficients are used to describe the formant structure of a speech waveform. Many factors contribute to the frame-to-frame fluctuation of these parameters. These variations adversely affect the performance of the LP quantizer and the quality of the synthesized speech. For voiced speech, efficient coding of the pitch pulses at the output of the inverse formant filter relies on the similarity of successive pitch waveforms. The performance of this coding stage is also jeopardized by LP variations. In this paper, we propose a new method which smoothes the evolution of the LP parameters. Our algorithm is based on matching the output of the formant predictor to a target signal constructed using smoothed pitch pulses. With this approach we have successfully reduced the frame-to-frame variation of LP coefficients, while increasing the similarity of pitch pulses.

1. INTRODUCTION

The first step in coding the discrete speech samples is the elimination of the near sample redundancies by means of an FIR filter, known as the inverse formant filter or the short term predictor. The coefficients of this filter are calculated using a standard linear prediction analysis and are updated every 20–30 ms. In addition, coding the residual signal involves modeling the pitch pulses. In Code Excited Linear Predictive Coders (CELP) this is accomplished by means of an analysis-by-synthesis strategy where the best possible waveform is selected from an adaptive codebook containing past pitch pulses, whereas in Adaptive Predictor Coders (APC) a pitch predictor is used. The performance of this pulse coding stage depends on the similarity of the successive pitch waveforms at the output of the inverse formant filter.

The shortcomings of the standard linear prediction technique in modeling the vocal tract transfer function have been known for a long time. Deller [1] has shown that the output of the short term predictor is a phase altered version of the glottal signal. El-Jeroudi and Makhoul [2] have shown that for periodic signals the linear prediction analysis will fail to identify the true all-pole model parameters due to the aliasing in the autocorrelation domain. Moreover, since the short term prediction of the speech is performed on a frame-to-frame basis in asynchrony with the evolving speech waveform, artificial variations in the predictor parameters may be introduced by the location of the analysis window.

It is desirable for the formant filter parameters to evolve slowly, since their fluctuations may be accentuated under quantization, creating audible distortions at update instants. Also, variations of the LP coefficients lead to changes in the pitch pulse shape for the pulses located in adjacent frames. The naturalness of the synthesized speech and the coding efficiency is compromised by these fluctuations. The most common approach to reduce the effect of these variations is to interpolate LP coefficients at intervals of 5 to 10 ms. However, since this is accomplished independently of the evolving residual waveform, the differences in the pitch pulses shape (introduced by the changes in the LP coefficients) are not eliminated. In this paper, we propose a new method for deriving the coefficients of the formant filter which takes into account the evolution of the residual pulses. The smoothing then takes place jointly in the residual and in the LP domains.

2. TARGET MATCHING

Standard linear prediction computes the coefficients of the short term predictor by minimizing the energy of the output signal. The output signal contains the glottal excitation which consists of periodic pulses (voiced

This work was supported in part by the Fonds pour la Formation de Chercheurs et l'Aide à la Recherche (FCAR) and the Canadian Institute for Telecommunication Research(CITR)

speech) and/or noise-like signal (unvoiced speech). It

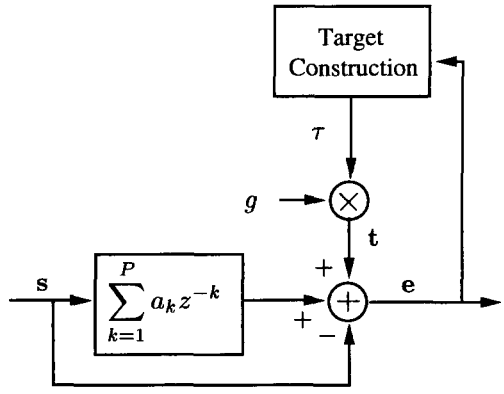

Figure 1: Target matching analysis block diagram.

has been shown [3] that the order in which the formant and pitch prediction take place affects their performance. When the formant filter precedes the pitch predictor then the latter's prediction gain is lower than if the order of the predictors is reversed. This implies that the formant predictor also participates in the task of pitch pulse modeling.

From the output of the standard LP filter, it is possible to form a target residual signal containing slowly evolving pulses. Our method then rederives the coefficients of the short term predictor so as to minimize the difference (MSE sense) between the filter output and this target. The new analysis filter (Fig. 1) is a Wiener filter which minimizes the error between the residual and a target waveform. Providing the excitation signal as a target relieves the short term predictor from the task of modeling the pulse shapes. Let $\mathbf{s}$, $\mathbf{S}$ and $\mathbf{t}$ be the speech frame, the data matrix and the target signal, respectively. The error is given by

$$\mathbf{E} = \mathbf{e}^T\mathbf{e} = (\mathbf{Sa} - \mathbf{s} + \mathbf{t})^T(\mathbf{Sa} - \mathbf{s} + \mathbf{t}) \qquad (1)$$

Solving $\nabla_{\mathbf{a}}\mathbf{E} = 0$ and $\nabla_g\mathbf{E} = 0$, leads to

$$(\mathbf{S}^T\mathbf{S})\mathbf{a} = \mathbf{S}^T(\mathbf{s} - \mathbf{t}) \qquad (2)$$

$$g = \frac{\mathbf{s}^T\tau - \tau^T\mathbf{P}_s\mathbf{s}}{\tau^T\tau - \tau^T\mathbf{P}_s\tau} \quad \text{where} \quad \mathbf{P}_s = \mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T \qquad (3)$$

Several points can be noted:

- Depending on the structure of matrix $\mathbf{S}$, $\mathbf{S}^T\mathbf{S}$ will correspond to the autocorrelation or the covariance matrix [4].
- The scale factor g is introduced to eliminate the influence of target gain on the residual shape. With properly scaled pulses in the target τ, g will be near unity.
- When the target signal is equal to zero, this method reduces to standard LP analysis.
- When the target signal is equal the original LP residual, then the second term in the right hand side of the Eq. (2) is zero by the orthogonality principal, therefore $\mathbf{a} = \mathbf{a}_{lpc}$. In this case, the target is perfectly matched. In the more general case

$$\mathbf{t} = \mathbf{e}_{lpc} + \xi \qquad (4)$$

$$\Delta\mathbf{a} = \mathbf{a} - \mathbf{a}_{lpc} = -(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\xi \qquad (5)$$

  Where the $\Delta\mathbf{a}$ is the correction to the standard linear prediction coefficients.
- $\mathbf{P}_s$ is the orthogonal projection matrix onto the columns of $\mathbf{S}$, and is equal to $\mathbf{U}^T\mathbf{U}$ [4] where $\mathbf{U}$ is the matrix of left singular vectors of $\mathbf{S}$.
- The resulting filter $\mathbf{a}$ is not guaranteed to be minimum phase.

The new residual signal obtained by filtering the speech with $\mathbf{a}$ is given by

$$\mathbf{x} = \mathbf{s} - \mathbf{P}_s(\mathbf{s} - \mathbf{t}) \qquad (6)$$

The optimality of the Wiener filter guarantees that $\|\mathbf{t} - \mathbf{x}\| \leq \|\mathbf{t} - \mathbf{e}_{lpc}\|$. This inequality implies that pulses in the modified residual evolve more slowly than those in the original LP residual. Pitch pulse coding efficiency is therefore increased.

3. TARGET CONSTRUCTION

The target signal should be as close as possible to the true excitation signal at the input of the vocal tract. To reduce the variations in the shape of adjacent residual pulses each target pulse is constructed considering past and possibly future pulses. The algorithm starts by smoothing LP coefficients corresponding to successive frames of speech. i.e.

$$\hat{\mathbf{a}}_{\mathrm{i}} = I(\mathbf{a}_{\mathrm{i}-1}, \mathbf{a}_{\mathrm{i}}) \qquad (7)$$

where I is an interpolation function operating in the LP coding domain, $\mathbf{a}_{\mathrm{i}}$ and $\mathbf{a}_{\mathrm{i}-1}$ are the predictor coefficients of the current and the previous frames, respectively. The speech signal is then filtered using this new set of parameters to form the residual signal $\mathbf{r}$ which serves as the input to the target construction algorithm. Our approach is based on the assumption that each pulse $\mathbf{y}$, is composed of two orthogonal components [5], the underlying pulse $\mathbf{v}$ which is nearly constant for the adjacent pulses. The innovation component $\mathbf{u}$ models variations due to changes in the underlying pulse and due to changes in the LP parameters.

$$\mathbf{y} = \beta\mathbf{v} + \alpha\mathbf{u} \qquad (8)$$

The vectors $\mathbf{y}$, $\mathbf{v}$, and $\mathbf{u}$ are normalized to unit energy, $\beta = \mathbf{v}^T\mathbf{y}$, $\alpha = \mathbf{u}^T\mathbf{y}$. Let $\mathbf{x}_0 \ldots \mathbf{x}_{L-1}$ be the pulses in $\mathbf{r}$. To construct the target pulse for $\mathbf{x}_l$, after normalization and appropriate alignment, we form the matrix

$$\mathbf{Y}_l = \left[\mathbf{v}_{-n_1+l} \;\cdots\; \mathbf{v}_{l-1} \;\; \mathbf{y}_l \;\cdots\; \mathbf{y}_{\min(n_2+l,L-1)}\right] \quad (9)$$

The target pulse for $\mathbf{y}_l$ is computed by minimizing the matching error,

$$\begin{aligned} \mathbf{v}_l &= \arg\min_{\|\mathbf{v}\|=1} \sum_i {\alpha_i}^2, \quad \alpha_i = \mathbf{u}^T\mathbf{y}_i \\ &= \arg\max_{\|\mathbf{v}\|=1} \sum_i {\beta_i}^2, \quad \beta_i = \mathbf{v}^T\mathbf{y}_i \\ &= \arg\max_{\|\mathbf{v}\|=1} \|{\mathbf{Y}_l}^T\mathbf{v}\| \end{aligned} \quad (10)$$

The vector $\mathbf{v}_l$ is the first right singular vector of $\mathbf{Y}_l$. The target frame τ is obtained by replacing $\mathbf{x}_l$ with the scaled and aligned version of $\mathbf{v}_l$. The following figure illustrates three frames of the standard LP residual and the corresponding target signal. Compared to the original residual, the smooth evolution in the target pulses shape is clearly noticeable.

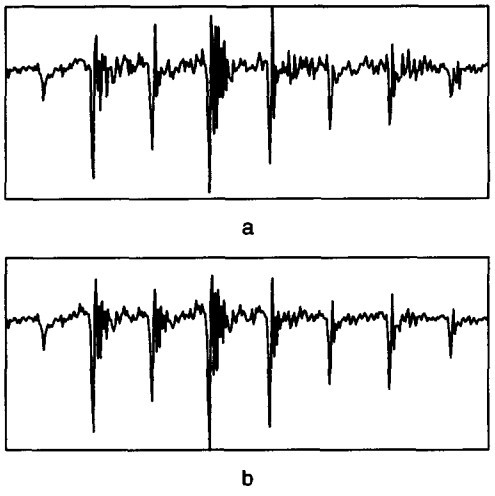

Figure 2: (a) Standard LP residual. (b) Target signal.

4. SPECTRAL SMOOTHNESS AND STABILITY

As a measure of smoothness in the evolution of the filter coefficients, we use the norm-1 of the difference vector between consecutive set of coefficients in the LP coding domain. Since the target pulses that are being matched are extracted from $\mathbf{r}$, the new LP parameters tend to be closer than the original LP coefficients to those of the previous frame. In the predictor coefficients domain, a sufficient condition for

$$\|\mathbf{a}_i - \mathbf{a}_{i-1}\|_1 \;<\; \|\mathbf{a}_{lpc} - \mathbf{a}_{i-1}\|_1 \quad (11)$$

is that every element of $\Delta\mathbf{a}$ (Eq. 5) respects the following:

$$0 \;\leq |\Delta a| \;\leq\; 2|a_{i-1} - a_{lpc}| \quad (12)$$

In situations where the inequality (11) is not respected one can reduce the number of the columns of $\mathbf{Y}$ to make the target more similar to the residual; as a result $\Delta\mathbf{a}$ will decrease. Another solution is to replace ξ by $\mu\xi$, where $\mu < 1$ (Eq. 4). Reducing gradually the weight factor μ causes $\mathbf{a}$ to approach $\mathbf{a}_{lpc}$ until the inequality (11) is satisfied. We adopt this last approach because of its low computational requirement. The same solutions may be applied in the situations where the new LP filter is not minimum phase.

5. EXPERIMENTS

High pitch female speech was first sampled at 8kHz. Standard linear prediction coefficients were calculated every 20 ms, using a 30 ms Hamming analysis window for the autocorrelation method. The resulting parameters were smoothed with respect to the previous frame (7). To filter the input speech, these parameters were held constant for 40 samples and linearly interpolated (in either LSF or PARCOR domain) between adjacent frames. Pitch pulse extraction took place on the output residual $\mathbf{r}$ using an independent pulse detection[2] algorithm. Each target pulse is constructed considering the three previous and the two future pulses, i.e. $n_2 - n_1 + 1 = 6$ (Eq. 9).

If the matrix $\mathbf{S}^T\mathbf{S}$ is desired to be Toeplitz, then the target signal $\mathbf{t}$ should contain the edge effect introduced by windowing the input speech. However, these edge values depend on the filter $\mathbf{a}$ for which the system (1) is being solved. To sidestep this problem, these edge values were estimated iteratively. The first and last P samples[3] of the target were replaced by those of the LP residual. System (2) is then solved for $\mathbf{a}$. For the second step, the edge values of target are updated by those of the residual at the the output of this new filter. We then iterate for $\mathbf{a}$. Experiments indicate that the above iteration converges rapidly.

To further improve on the similarity of the successive pitch pulses and reduce the frame-to-frame variation of the LP coefficients, the described target construction and matching algorithm was also applied in an iterative fashion. At the k-th step, once the stability and the smoothness conditions are satisfied, the new LP filter $\mathbf{a}^{(k)}$ is fed back into the target construction routine and a new smoothed residual signal $\mathbf{r}^{(k+1)}$ is formed. The new target is constructed based on the pulses extracted from this new waveform. The resulting filter $\mathbf{a}^{(k+1)}$ is accepted if the smoothness in LP

[2]The information regarding pulse locations is in part available in many of the new generation coders [6] [7].

[3]P is the order of the predictor. In this paper $P = 10$.

parameters is improved, and if the matched residual pulses, $\overline{\mathbf{v}}$, are more similar than in the previous step:

$$\begin{aligned} \|\mathbf{a}^{(k+1)} - \mathbf{a}_{i-1}\|_1 &< \|\mathbf{a}^{(k)} - \mathbf{a}_{i-1}\|_1 \\ \sum_{l=1}^{L-1} (\overline{\mathbf{v}}_l^{(k+1)})^T (\overline{\mathbf{v}}_{l-1}^{(k+1)}) &< \sum_{l=1}^{L-1} (\overline{\mathbf{v}}_l^{(k)})^T (\overline{\mathbf{v}}_{l-1}^{(k)}) \end{aligned} \quad (13)$$

For the performance measure, we use the prediction gain (ratio of the signal energy, in dB, at the input to the signal energy at the output of the formant predictor). The similarity between successive pitch pulses may be evaluated by predicting one pulse from another; this is in effect a pitch predictor. The prediction gain for this predictor will then measure the similarity of adjacent pitch pulses. To monitor the smoothness in the formant filter coefficients, we measure the average of the norm-1 of the LP parameters difference vector in LSF (ω) or predictor coefficients ($\mathbf{a}$) domains:

$$\overline{\|\Delta\omega\|}_1 = \sum_{i=1}^{N-1} \|\omega_{i+1} - \omega_i\|_1/(N-1) \quad (14)$$

Where N is the total number of frames. Since the standard LP filter is optimal in the MSE sense, a decrease in the short term prediction gain is expected. However, this loss is more than compensated for by an increase in the pitch pulse similarities as measured by the pitch prediction gain.

Matching Method	Prediction Gain (dB) Formant	Pitch	$\overline{\|\Delta\omega\|}_1$
LP	12.6	5.8	0.75
TM	12.4	6.4	0.64
ITM	11.9	6.8	0.60

Table 1: Autocorrelation method, LP: Standard Linear Prediction, TM: Target matched, ITM: Iterative TM.

Matching Method	Prediction Gain (dB) Formant	Pitch	$\overline{\|\Delta\mathbf{a}\|}_1$
LP	12.6	5.9	2.66
TM	12.2	6.4	2.11
ITM	12.0	6.5	2.00

Table 2: Covariance method.

Optimizing the LP filter according to the target signal results in only a small loss in the formant prediction gain. The benefit of the proposed analysis method is an increase in the smoothness of the filter dynamics. Consequently, the successive pulses in voiced regions are more similar, and the pitch prediction gain has also increased. The price for the higher performance of the iterative approach is the larger reduction in the formant predictor gain and the extra computation.

For the covariance method, the matching process actually reduces the number frames with unstable LP parameters. For the autocorrelation method, although the minimum phase property of $\mathbf{a}$ is not guaranteed, all the resulting LP synthesis filters were stable for the tested speech segments.

6. CONCLUSION

In this paper, we have presented an alternative method to perform the linear prediction analysis of a speech signal. The inverse formant filter is replaced by a Wiener filter with the target signal containing slowly evolving pulses. Experiments show that the frame-to-frame variation of LP coefficients is reduced and the matched residual pitch pulses evolve more slowly with time. The price for these gains in coding efficiency is paid in terms of the amount of the computation required to construct the target and to derive the filter with the best overall performance.

7. ACKNOWLEDGMENT

We would like to thank Mr. Jacek Stachurski for providing us with his robust pitch pulse detection program.

8. REFERENCES

[1] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals.* MacMillan, 1994.

[2] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Processing*, vol. 39, pp. 411–423, Feb. 1991.

[3] P. Kabal and R. P. Ramachandran, "Joint optimization of linear predictors in speech coders," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 37, pp. 642–650, May 1989.

[4] C. W. Therrien, *Discrete Random Signals And Statistical Signal Processing.* Prentice Hall, 1992.

[5] J. Stachurski, "A pitch pulse evolution model for the linear predictive coding of speech," *Ph.D. Proposal, Dept. Electrical Eng. McGill University, (unpublished)*, pp. 20–28, 1994.

[6] W. B. Kleijn, P. Kroon, and D. Nahumi, "The rcelp speech-coding algorithm," *European Trans. on Telecom. and Related Technologies*, vol. 5, pp. 573–582, Sept. 1994.

[7] 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.