Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (New York, NY), pp. 315-318, April 1988

JOINT SOLUTIONS FOR FORMANT AND PITCH PREDICTORS IN SPEECH PROCESSING

Peter Kabal[1 2] and Ravi P. Ramachandran[1]

[1]Electrical Engineering
McGill University
Montreal. Quebec H3A 2A7

[2]INRS-Télécommunications
Université du Québec
Verdun, Quebec H3E 1H6

Abstract

The statistical correlation between speech samples motivates the use of prediction error filters that can remove this redundancy. Formant filters act to remove the near-sample redundancies while pitch filters remove the far-sample redundancies for samples separated by the pitch period. The use of these filters is especially beneficial in low bit rate speech coding. The coefficients of these predictors are usually determined sequentially. This paper discusses jointly optimized solutions for the two predictors. The filters are implemented in transversal form with the formant filter always preceding the pitch filter (F-P cascade). The jointly optimized methods are further subdivided into a combined solution and an iterative sequential solution. Both yield higher prediction gains than their conventional sequential determined counterpart. Furthermore, a practical version of the iterated sequential approach in which the filters are constrained to be minimum phase generates decoded speech of high perceptual quality when applied in a coding environment.

1. Introduction

A primary concern in speech analysis is the effective removal of both near and distant sample based redundancies. This is often performed by two nonrecursive prediction error filters each devoted to the elimination of a specific type of redundancy. Formant predictors remove near sample correlations while pitch predictors remove distant sample correlations. The incoming speech signal is processed by these two filters with the coefficients chosen to minimize the mean-square value of the output residual.

Usually, a cascade connection of the two prediction error filters is implemented with the filter coefficients being calculated sequentially. The coefficients of the first predictor are determined and used to generate an intermediate residual. Then the coefficients of the second predictor are determined and used to generate the final residual. Previous studies [1][2] have shown that placing the formant predictor before the pitch predictor (F-P cascade) results in a residual signal of lower energy than the complementary connection.

In this paper, we restrict ourselves to an F-P cascade but generate jointly optimal solutions for the formant and pitch predictors. The combined optimization which minimizes the mean-square value of the final residual is developed in Section 3. This section continues by introducing the iterated sequential approach and a practical implementation of it which forces the prediction error filters to be minimum phase and which makes use of an efficient method of computing the pitch lag. The practical implementation of the iterated sequential approach is evaluated as part of a Code-Excited Linear Predictive (CELP) coder.

2. General Analysis of Linear Predictors

Figure 1 depicts a general analysis model for a linear predictor with arbitrary delays M_k. Both the input and error signals are windowed. By minimizing the squared error sum $\varepsilon^2 = \sum_{n=-\infty}^{\infty} e_w^2(n)$, a linear system of equations [2] ($\mathbf{\Phi c} = \boldsymbol{\alpha}$) are obtained. The entries of the positive definite symmetric matrix $\mathbf{\Phi}$ are given by

$$\phi(i,j) = \sum_{n=-\infty}^{\infty} w_e^2(n)\, x_w(n - M_i)\, x_w(n - M_j) \; . \qquad (1)$$

The vectors $\mathbf{c}$ and $\boldsymbol{\alpha}$ are given by

$$\mathbf{c}^T = [c_1, c_2, \ldots, c_L] \qquad \text{and} \qquad \boldsymbol{\alpha}^T = [\phi(0,M_1), \ldots, \phi(0,M_L)] \; . \qquad (2)$$

This general formulation subsumes the autocorrelation and covariance approaches. Note that the autocorrelation method is generally not suitable for predictors with tap delays which are an appreciable fraction of the frame size. Moreover, the autocorrelation method does not ensure that such prediction error filters are minimum phase [2]. Hence, this method is not appropriate for pitch analysis.

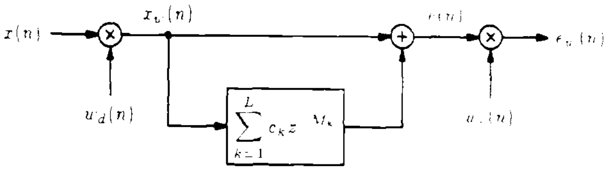

Fig. 1 Analysis model for transversal predictors

The general linear predictor described above subsumes both formant and pitch predictors. A formant predictor $F(z)$ has continuous support in that prediction is based on N_f previous samples (usually $8 \le N_f \le 16$).

$$F(z) = \sum_{k=1}^{N_f} a_k z^{-k} \qquad (3)$$

For this case, the autocorrelation method guarantees that $1 - F(z)$ is minimum phase [3]. Although the covariance method does not necessarily give a minimum phase solution, an alternative modified covariance method based on residual energy ratios [4][5] does assure minimum phase.

The pitch predictor has a small number of taps, N_p clustered around a large delay which corresponds to the estimated pitch period M. Its system function is

$$P(z) = \begin{cases} \beta_1 z^{-M} & \text{1 tap} \\ \beta_1 z^{-M} - \beta_2 z^{-(M-1)} & \text{2 tap} \\ \beta_1 z^{-M} - \beta_2 z^{-(M-1)} - \beta_3 z^{-(M-2)} & \text{3 tap} \end{cases} \qquad (4)$$

Methods to choose the pitch lag are discussed in [2]. Generally, a covariance analysis determines the coefficients β_i. Although minimum phase is not assured, a fixup procedure developed in [6] can be utilized to achieve minimum phase with negligible loss in performance.

3. Joint Solutions

Joint solutions for the formant and pitch predictors are accomplished through combined optimization and an iterative sequential approach. By minimizing the mean-square value of the final residual directly, the formant predictor ignores pitch pulses in the intermediate residual which are subsequently removed by the pitch predictor. Hence, each predictor concentrates on eliminating the type of sample correlations that it is best suited to handle.

3.1 Combined Optimization

A combined optimization can be either performed in one-shot or iteratively depending upon the relative values of the frame size N and the pitch lag M. In particular, if $M \ge N$, the optimization can be carried out in one-shot since only the formant predicted residual from previous frames is required as the input to the pitch filter. The F-P cascade shown in Fig. 2 is the configuration used for the analysis. Note that the analysis is performed by decoupling the input to the pitch filter from the output of the formant filter. The coefficients are computed by minimizing the rectangularly windowed mean-square error $E = \sum_{n}^{N-1} e^2(n)$ to get a system of linear equations $\mathbf{\Phi c} = \boldsymbol{\alpha}$, where

$$\mathbf{c}^T = [a_1, \ldots, a_{N_f}, \beta_1, \ldots, \beta_{N_p}] ,$$

$$\mathbf{\Phi} = \sum_{n=0}^{N-1} \mathbf{u}^{(n)} \mathbf{u}^{(n)T} \qquad \text{and} \qquad \boldsymbol{\alpha} = \sum_{n=0}^{N-1} s(n) \mathbf{u}^{(n)} \tag{5}$$

and the data vector is defined by

$$\mathbf{u}^{(n)T} = [s(n-1), \ldots, s(n-N_f), d(n-M), \ldots, d(n-M-N_p-1)] . \tag{6}$$

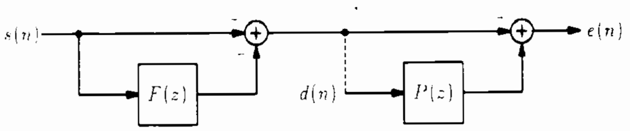

Fig. 2 F-P cascade configuration

The limitation that $M \geq N$ can be circumvented by using an iterative approach to calculate the combined solution. This scheme uses the intermediate residual signal $d(n)$ from a previous iteration to solve the system of equations. For the first iteration, a conventional formant filter formulation is used (i.e. the pitch coefficients are set to zero). The formant filter is used to filter the input signal to produce the intermediate residual signal $d(n)$. The next iteration uses a combined optimization, but based on $d(n)$ from the previous iteration. The updated formant filter is used to generate an updated intermediate residual signal. The iterations continue using the combined solution in this manner until the error no longer decreases significantly. Note that this algorithm does not guarantee that the overall error decreases monotonically.

3.2 Iterative Sequential Approach

An alternate joint solution is built around a sequential optimization of an F-P cascade. During the first iteration, the formant predictor coefficients are chosen to minimize the intermediate residual energy. The pitch predictor coefficients are then found using this intermediate residual signal. At this point, the filter coefficients are those that would be used when a sequential solution is applied to a conventional F-P cascade. During subsequent iterations, the formant filter is reoptimized given the previously determined pitch predictor coefficients. Also, the pitch filter is recalculated based on the newly formed intermediate residual. The overall residual energy monotonically decreases with each step to a local minimum.

The equations to reoptimize the formant filter taking into account the pitch filter can be viewed as minimizing a weighted mean-square error criterion. The analysis model is shown in Fig. 3. For the case at hand, the weighting filter $H(z)$ is the pitch prediction error filter. An additional input signal $e_0(n)$ captures the initial conditions of the weighting filter. In our case, $e_0(n)$ is the output of the pitch filter due to inputs from previous analysis frames and cannot be neglected for normal frame sizes. There is an additional window applied to the intermediate residual signal. Minimization of the weighted error energy leads to a set of linear equations $\mathbf{\Phi c} = \boldsymbol{\alpha}$, where $\mathbf{c}$ is the vector of formant predictor coefficients. The elements of $\mathbf{\Phi}$ are given by

$$\phi(i,j) = \sum_{n=-\infty}^{\infty} w_e^2(n)\, s^{(i)}(n)\, s^{(j)}(n) , \tag{7}$$

where

$$s^{(i)}(n) = \sum_{m=-\infty}^{\infty} w_r(m) h(n-m) s_u(m-i) . \tag{8}$$

The elements of $\boldsymbol{\alpha}$ are

$$\alpha_i = \sum_{n=-\infty}^{\infty} w_e^2(n) \, [s^{(0)}(n) - e_0(n)] \, s^{(i)}(n) . \tag{9}$$

The elements of $\mathbf{c}$ are the formant predictor coefficients. Consider a covariance analysis with rectangular residual and error windows (nonzero for n ranging from 0 to $N-1$) and a causal weighting filter. The form of the signal $s^{(i)}(n)$ can then be simplified.

$$s^{(i)}(n) = \sum_{m=0}^{n} h(n-m) s(m-i) , \quad 0 \leq n < N . \tag{10}$$

A practical implementation of the iterated sequential method in which both the formant and pitch filters are minimum phase is now described. The filters will be optimized together as long as the minimum phase property is not sacrificed. In addition, a practical method to determine the pitch lag is used.

For the first iteration, a covariance approach is used to determine a formant filter. If this filter is not minimum phase, as determined from the magnitudes of the corresponding reflection coefficients, new coefficients are found using a modified covariance approach which guarantees a minimum phase solution. The input signal is filtered to produce the intermediate residual signal. The pitch lag is determined using the method described in [2]. The pitch filter coefficients are found using a covariance analysis. If the pitch filter is not minimum phase, it is stabilized by the algorithm developed in [6].

For subsequent iterations, the pitch lag is fixed at the value determined at the first iteration. Moreover, subsequent iterations allow for the formant filter to take into account the pitch filter using the analysis derived for the iterated sequential method. Again, a covariance analysis is used. If the formant filter is unstable, one reverts to a modified covariance analysis for the formant filter. Note that the modified covariance analysis does not take into account the effect of the pitch filter. The pitch filter is again determined using a covariance analysis and stabilized if necessary. These iterations continue until the overall gain no longer increases significantly.

4. Simulation Results

This section presents the experimental results obtained for the F-P cascade connection in which the coefficients are computed by the joint optimization algorithms described in this paper. The formant predictor has 10 coefficients. One, two and three tap pitch predictors are used. A total of six sentences were processed of which three were spoken by a male and three by a female. The sampling frequency is 8 kHz. The average prediction gains in dB are computed for each sentence in order to compare the different methods. Since the comparisons are essentially consistent for each sentence, the tables which follow present average values of the prediction gains in dB taken over the six sentences. The tables refer to the formant gain, meaning the ratio of the energy of the input to the formant filter to the energy of the output of the formant filter. A similar definition holds for the pitch gain.

4.1 Performance and Stability Issues for the Joint Solutions

Here, the joint solutions are compared to the conventional sequential approach. For this purpose, covariance analyses are also used for both the formant and pitch predictors in the conventional sequential approach. The value of the pitch lag M is chosen such that the prediction gain is maximized. This is accomplished by an exhaustive search over the allowable range. For frame sizes of 80 samples, the value of M lies between 20 and 120.

Table 1 presents the prediction gains for the two joint solutions. The combined solution uses the one-shot scheme for pitch lags larger than the frame size and uses the iterated approach for smaller pitch lags. Experiments show that an average of three iterations (0.02 dB prediction gain threshold) are needed to resolve the coefficients. The iterated sequential method gives larger prediction gains than the combined optimization method at the expense of a slightly larger average number of iterations (5 per frame). For the combined optimization scheme, the iterations are often terminated by a decrease in overall prediction gain (nonmonotonicity of decrease in error).

method	formant gain dB	pitch gain dB	overall gain dB
F-P sequential	16.2	4.4 5.6 6.1	20.6 21.8 22.3
F-P combined solution	15.3 14.1 14.2	6.4 9.2 9.5	21.7 23.3 23.7
F-P iterated sequential	14.8 14.0 14.1	7.2 9.5 9.8	22.0 23.5 23.9

Table 1 Prediction gains for the sequential and jointly optimized predictors (80 sample frames). Three numbers in an entry refer to 1, 2 and 3 tap pitch filters. A covariance formulation is used for all cases.

Table 2 gives figures for the percentage of frames with unstable synthesis filters for sequential optimization, iterated sequential optimization and combined optimization for 80 sample frames. The stability of the formant filter was determined by converting the predictor coefficients obtained to the equivalent reflection coefficients. The stability of the pitch filter was determined by the tight sufficient test given in [6]. The incidence of instability of the formant filters is much lower than that for the pitch filters. Iterated combined and iterated sequential optimization of an F-P cascade lead to more cases of instability than the conventional sequential solution.

method	formant % unstable	pitch % unstable
F-P sequential	4	6 26 25
F-P combined solution	8 6 8	12 29 35
F-P iterated sequential	7 6 8	11 30 34

Table 2 Stability of sequential and jointly optimized filters for 80 sample frames, covariance analysis. Three numbers in an entry refer to 1, 2 and 3 tap pitch filters.

4.2 Minimum Phase Solution

It is the incidence of instability in the iterated sequential approach that provides some of the motivation for the minimum phase solution described earlier. In this section, the question that is addressed is whether the increased prediction gain of the joint solutions comes only at the expense of unstable filters.

We maintain the same experimental conditions as before. However, a practical method of calculating M (see [2]) is utilized. The resulting prediction gains for a minimum phase iterated sequential solution are shown in Table 3. Also included in the table are the prediction gains if the process is stopped at the first iteration (a pure sequential approach). The results show that an increase in overall prediction gain is achieved by the iterative approach even when minimum phase filters are mandated. These results can also be compared to Table 1 to see a drop in overall prediction gain due to the use of stabilized filters and a practical method to find the pitch lag. However, the above two constraints do not substantially diminish the prediction gain. Moreover, the gain achieved by the constrained iterative approach still remains above that achieved by the unconstrained F-P sequential algorithm which admit both unstable formant and unstable pitch filters (see Table 1).

method	formant gain dB	pitch gain dB	overall gain dB
F-P sequential	16.2	4.4 5.3 5.8	20.6 21.5 22.0
F-P iterated sequential	15.6 15.2 15.2	6.0 7.5 7.9	21.6 22.7 23.1

Table 3 Prediction gains for minimum phase filters in an F-P cascade (80 sample frames). Three numbers in an entry refer to 1, 2 and 3 tap pitch filters.

Figure 4 shows plots of the prediction gains (3 tap pitch filter) for an utterance ("Thieves who rob friends deserve jail") spoken by a female. The largest increases in prediction gain due to joint optimization tend to be in the energetic voiced regions.

Figure 5 shows the formant and formant pitch predicted residuals for a voiced section of the same utterance. The iterated sequential method gives a formant residual with larger, but more regular pitch pulses than the sequential method. The efficacy of the pitch filter is enhanced by these regular pitch pulses. The pitch filter can then form a final residual having a smaller magnitude and containing fewer spurious bursts than the conventional sequential approach.

The conventional sequential and iterated sequential approaches (both using stabilized filters and the practical method to choose the pitch lag) were implemented as part of a CELP coder. The analysis frame size remains at 80 samples. A 10th order formant predictor and a 3 tap pitch predictor were used. The coefficient values determined by the analysis stage are used for the corresponding synthesis filters. Forty sample blocks of the residual were compared to a dictionary of 1024 waveforms consisting of Gaussian random numbers with unit variance. The codeword that represents the residual is the one that achieves the least weighted error. The error is weighted by a filter $W(z) = (1 - F(z))/(1 - F(z/\alpha))$. This filter deemphasizes the frequencies which contribute less to perceptual error and emphasizes the frequencies which contribute more to perceptual error [7]. The noise weighting factor α equals 0.8.

Perceptual tests reveal that the iterated sequential algorithm leads to resynthesized speech that is more natural sounding than the decoded speech resulting from the sequential approach. Also, this speech sounds less warbled than its conventional counterpart. Although the perceived improvement varies from sentence to sentence, this phenomenon is clearly evident in the sentence "Thieves who rob friends deserve jail" and can be attributed to the fact that the residual as depicted in Fig. 5 has fewer spurious peaks when the iterated sequential method is used. The Gaussian codewords provide a better match to the residual when it has less energetic pitch pulses. Consequently, the weighted error is smaller and the decoded speech sounds more like the original.

5. Summary and Conclusions

This paper has introduced formulations which allow for the joint optimization of the formant and pitch predictors. The final residual energy can be directly minimized by a cascade of jointly optimized filters. This has a significant advantage in terms of increased prediction gain.

The limitations of the one-shot combined solution are that it is applicable only when the pitch lag exceeds the frame length and that it admits both unstable formant and pitch synthesis filters. However, an iterated combined scheme can be used to get around the frame length constraint. An iterated sequential solution achieves slightly more prediction gain than the iterated combined solution. Furthermore, the minimum phase constraint can be satisfied by backing off to the modified covariance method for the formant filter in cases of non-minimum phase filters and by stabilizing the pitch filter. The advantages of the iterated sequential method over the conventional sequential method lie not only in achieving a higher prediction gain but also in resynthesizing speech of better perceptual quality.

References

1. J. L. Flanagan, M. R. Schroeder, B. S. Atal, R. E. Crochiere, N. S. Jayant and J. M. Tribolet, "Speech coding", *IEEE Trans. Commun.*, vol. COM-27, pp. 710–736, April 1979.
2. R. P. Ramachandran and P. Kabal, "Pitch prediction filters in speech coding", to appear in *IEEE Trans. Acoust., Speech, Signal Processing*.
3. S. W. Lang and J. H. McClellan, "A simple proof of stability for all-pole linear prediction models", *Proc. IEEE*, vol. 67, pp. 860–861, May 1979.
4. B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 247–254, June 1979.
5. B. W. Dickinson, "Autoregressive estimation using residual energy ratios", *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 503–506, July 1978.
6. R. P. Ramachandran and P. Kabal, "Stability and performance analysis of pitch filters in speech coders", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 937–946, July 1987.
7. M. R. Schroeder and B. S. Atal, "Code-Excited Linear Prediction (CELP): High-quality speech at very low bit rates", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tampa, Florida, pp. 25.1.1–25.1.4, March 1985.

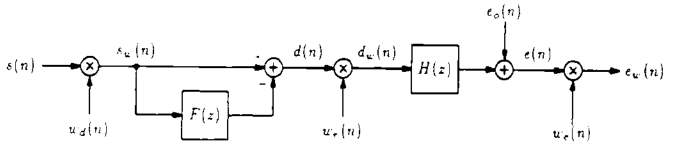

Fig. 3 Analysis model for a prediction filter with error weighting

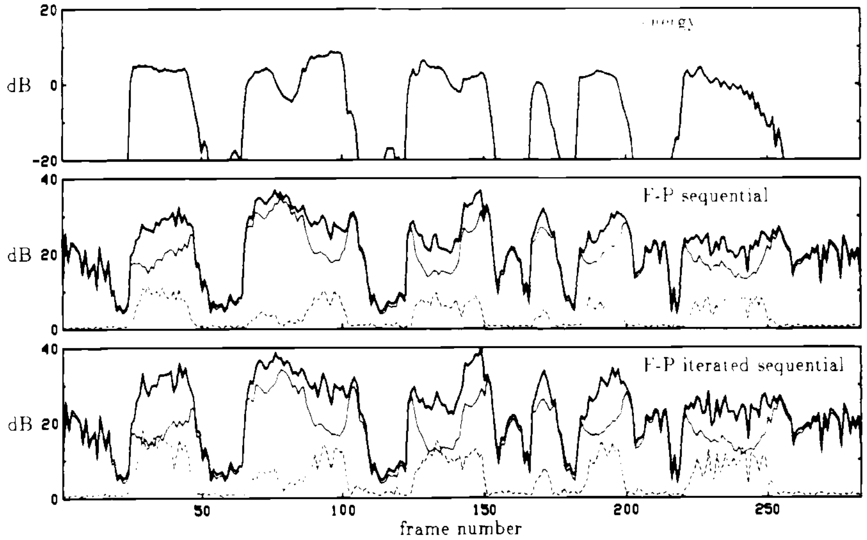

Fig. 4 Prediction gains (3 tap pitch filter), frame by frame. In each of the two lower plots, the upper trace is the overall prediction gain, the middle light trace is the formant prediction gain and the lower dashed trace is the pitch prediction gain.

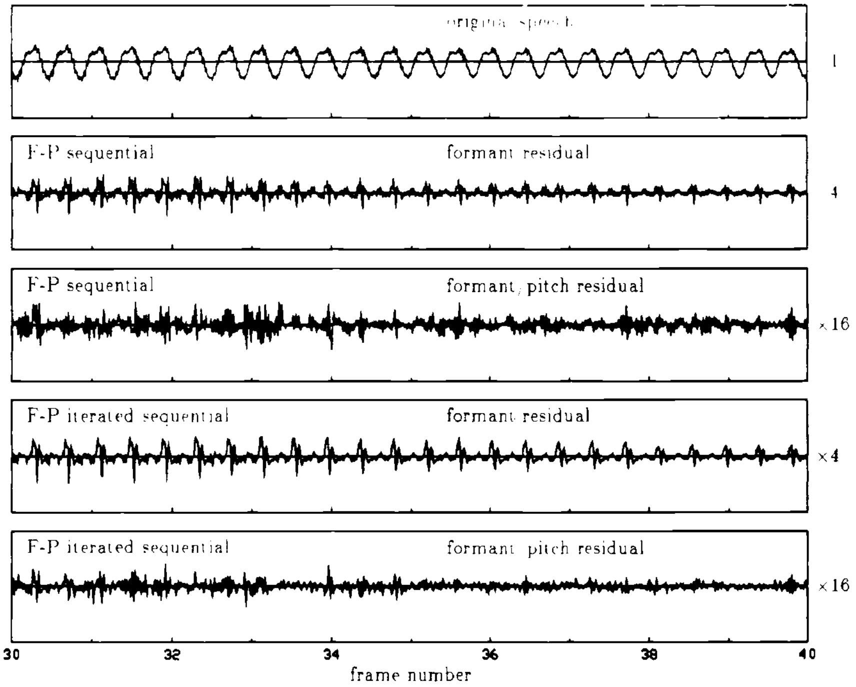

Fig. 5 Prediction residuals (3 tap pitch filter) for a section of voiced speech. Magnification factors are noted beside each plot.