Joint Optimization of Linear Predictors in Speech Coders

PETER KABAL, MEMBER, IEEE, AND RAVI P. RAMACHANDRAN

Abstract—**Low bit rate speech coders often employ both formant and pitch predictors to remove near-sample and distant-sample redundancies in the speech signal. The coefficients of these predictors are usually determined for one prediction filter and then for the other (a sequential solution). This paper deals with formant and pitch predictors which are jointly optimized. The first configuration considered is a combination prediction error filter (in either a transversal or a lattice form) that performs the functions of both a formant and a pitch filter. Although a transversal combination filter outperforms the conventional F-P (formant followed by pitch) sequential solution, the combination filter exhibits a high incidence of nonminimum phase filters. For an F-P cascade connection, combined solutions and iterated sequential solutions are found. They yield higher prediction gains than the conventional F-P sequential solution. Furthermore, a practical implementation of the iterated sequential solution is developed such that both the formant and pitch filters are minimum phase. This implementation leads to decoded speech of higher perceptual quality than the conventional sequential solution.**

I. Introduction

In low bit rate predictive coding of speech, two nonrecursive prediction error filters are often used to process the input signal before coding. The prediction operations are motivated by the fact that the input speech exhibits a high degree of intersample correlation. These correlations occur between adjacent samples (near-sample redundancy) and for voiced speech, between samples separated by the pitch period (far-sample redundancy). Near-sample redundancies can be attributed to the filtering action of the vocal tract. The resonances of the vocal tract correspond to the formant frequencies in speech. Far-sample redundancies can be attributed to the pitch excitation of voiced speech. Two filters, the formant and pitch predictors, are used to remove the near-sample and far-sample redundancies, respectively. The resulting prediction residual signal is of smaller amplitude and can be coded more efficiently than the original speech waveform. The predictor coefficients are adapted by updating them at fixed intervals to follow the time-varying correlation of the speech signal. An example of a system which uses the two predictor arrangement is an Adaptive Predictive Coder (APC).

Manuscript received September 4, 1987; revised September 8, 1988. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

P. Kabal is with the Department of Electrical Engineering, McGill University, Montreal, P.Q., Canada, H3A 2A7 and INRS-Telecommunications, Université du Québec, Verdun, P.Q., Canada, H3E 1H6.

R. P. Ramachandran is with the Department of Electrical Engineering, McGill University, Montreal, P.Q., Canada, H3A 2A7.

IEEE Log Number 8926668.

In conventional APC, the predictors are placed in a feedback loop around the quantizer. The quantization occurs sample-by-sample. With this configuration, it can be shown that the quantization noise is not only the difference between the residual and its quantized value but also the difference between the original speech signal and its reconstructed value. The perceptual distortion of the output speech can be reduced by adding a noise shaping filter which redistributes the quantization noise spectrum [1], [2]. The noise shaping filter increases the noise energy in the formant regions but decreases the noise power at frequencies in which the energy level is low. Its system function is often chosen to be a bandwidth expanded version of the transfer function of the formant predictor.

An alternate APC configuration places the predictors in an open-loop format and includes a noise shaping filter [3] as depicted in Fig. 1. The quantization is again accomplished sample-by-sample. Code-Excited Linear Prediction (CELP) [4] combines an open-loop arrangement for the predictors with vector quantization. Vector quantization is implemented by searching a given repertoire of waveforms for a candidate waveform that best represents the residual in a weighted mean-square sense. The weighting is employed to accomplish noise shaping.

The synthesis phase is similar in APC and CELP. In both cases, an excitation signal (the coded residual or the selected codeword after scaling) is passed through a pitch synthesis and a formant synthesis filter to produce the decoded speech. The synthesis operation can be viewed in the frequency domain as first inserting the periodic structure due to pitch and then inserting the spectral envelope (formant structure).

A previous paper has considered the cascade connection of a formant and pitch predictor [5]. In that paper, the coefficients were determined using a conventional sequential approach. For a sequential solution, the coefficients of the first predictor are determined from the input speech, and the coefficients of the second predictor are determined from the intermediate residual formed by the filtering action of the first predictor. The objective of this paper is to consider *combination* configurations and *joint* solutions for the formant and pitch filter coefficients. We present new algorithms for the joint optimization which give improved performance over standard techniques. The minimum phase property of the filters is also considered. A minimum phase prediction error filter at the analysis phase guarantees a stable synthesis filter. This is a signif-

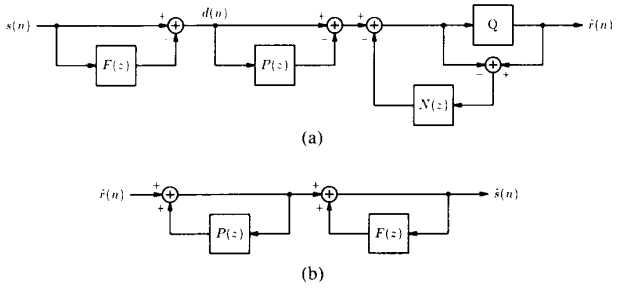

Fig. 1. Block diagram of an APC coder with noise feedback. (a) Analysis phase. (b) Synthesis phase.

icant issue since, if the synthesis filters are unstable, the quantization noise is accentuated and causes undesirable perceptual distortion in the output speech [6]. The final configuration considered constrains the solutions to be minimum phase. In addition, that system uses a simplified method to choose an appropriate pitch lag. This practical approach retains the gains due to joint optimization and gives real improvements in speech quality in a coding environment.

II. Formulation for Linear Predictors

Fig. 2 shows a general analysis model for a linear predictor with arbitrary delays M_k. Windows are applied to both the input and error signals. The aim of the analysis is to minimize the squared error sum $\epsilon^2 = \Sigma_{n=-\infty}^{\infty} e_w^2(n)$. This leads to a linear system of equations [5] which can be written in matrix form ($\Phi \boldsymbol{c} = \boldsymbol{\alpha}$) as

$$\begin{bmatrix} \phi(M_1, M_1) & \phi(M_1, M_2) & \cdots & \phi(M_1, M_L) \\ \phi(M_2, M_1) & \phi(M_2, M_2) & \cdots & \phi(M_2, M_L) \\ \vdots & \vdots & & \vdots \\ \phi(M_L, M_1) & \phi(M_L, M_2) & \cdots & \phi(M_L, M_L) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_L \end{bmatrix} = \begin{bmatrix} \phi(0, M_1) \\ \phi(0, M_2) \\ \vdots \\ \phi(0, M_L) \end{bmatrix}, \quad (1)$$

where the correlation entries are given by

$$\phi(i, j) = \sum_{n=-\infty}^{\infty} w_e^2(n) x_w(n-i) x_w(n-j). \quad (2)$$

Applying a window only to the input signal results in the autocorrelation method. The covariance formulation results when only the error signal $e(n)$ is windowed. Note that the autocorrelation method is not suitable for predictors with delays which are an appreciable fraction of the frame size [5]. Indeed, the autocorrelation method does not even guarantee minimum phase filters for general delays. These considerations rule out the use of the autocorrelation method for pitch analysis.

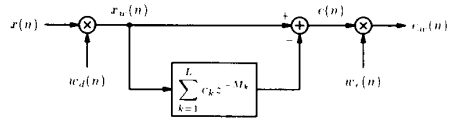

Fig. 2. Analysis model for linear predictors.

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 $8 \le N_f \le 16$),

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

For this case, the autocorrelation method guarantees that $1 - F(z)$ is minimum phase [7]. Although the covariance method does not necessarily give a minimum phase solution, an alternative approach called the modified covariance method is based on residual energy ratios [1], [8] and assures minimum phase.

The pitch predictor has a small number of taps, N_p. The delays associated with these taps are bunched around a value which corresponds to the estimated pitch period in samples. Its system function is

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

The pitch lag M is usually updated along with the coefficients. Methods to choose the pitch lag are discussed in [5]. The formant and pitch synthesis filters have transfer functions $H_F(z) = 1/(1 - F(z))$ and $H_P(z) = 1/(1 - P(z))$, respectively.

Conventionally, the formant and pitch predictors are connected in cascade. Furthermore, the determination of the predictor coefficients proceeds in two steps. The coefficients of the first predictor are chosen to minimize the energy of the intermediate residual signal. The filter coefficients so determined are then used to form the intermediate residual signal. The coefficients of the second predictor are computed with the aim of minimizing the energy of the final residual signal given the intermediate residual as the input signal. The predictors may be cascaded in either order. The sequential solution gives different sets of formant and pitch coefficients for the two orderings. In addition, the fact that the filters are time varying affects the initial conditions at frame boundaries. Hence, the different orderings of the predictors lead to different output residuals. Experiments show that having the formant filter precede the pitch filter (F-P cascade) renders a higher prediction gain than the reverse ordering (P-F arrangement) [5]. The upcoming experimental results are based on prediction gain as the performance measure. The prediction gain is the average energy of the input signal to the pre-

dictor divided by the average energy of the prediction error. The prediction gain is an appropriate measure since it assesses the extent to which redundancies are removed by the predictor.

III. Joint Formant/Pitch Solutions

Joint optimization of the formant and pitch predictors is accomplished in two different ways. First, the two predictors are combined into a single combination filter. Second, the predictors are connected as an F-P cascade with jointly optimized coefficients.

A. Combination Implementation

One method of achieving joint optimization is in a combination implementation in which the formant and pitch predictors are connected in parallel. The filter has two sets of grouped delays corresponding to its formant and pitch portions. The first group of delays ranges from 1 to N_f, while the second group ranges from M to $M + N_p - 1$. This combined filter can be viewed as a single prediction error filter $1 - F(z) - P(z)$ that processes the speech signal. A combination implementation results in a prediction error filter that has a different impulse response than the overall prediction error filter arising from a cascade connection. The overall impulse response of the combination filter has fewer nonzero coefficients ($N_f + N_p + 1$ for transversal implementation) than the cascade connection ($2N_f + N_p + 1$), although both have only $N_f + N_p$ degrees of freedom. A lattice implementation for a combination filter is also considered.

In a combination implementation, the coefficients of the transversal filter $1 - F(z) - P(z)$ are determined by placing a rectangular window on the error signal and using the system of equations given in (1) (covariance method). The coefficient vector $\boldsymbol{c}$ is given by $\boldsymbol{c}^T = [a_1, \cdots, a_{N_f}, \beta_1, \cdots, \beta_{N_p}]$. An inherent advantage of this configuration is that the formant and pitch predictor coefficients are jointly optimized. However, the combination formulation involves solving a larger system of equations ($N_f + N_p$ by $N_f + N_p$) than if the conventional sequential optimization is used. Given a covariance analysis, the main disadvantage is that even the formant part of the synthesis filter can be unstable. Although methods are available to obtain stable formant and pitch synthesis filters individually, the corresponding combination synthesis filter is not easily stabilizable.

A general all-zero lattice filter is shown in Fig. 3. A lattice structured combination filter can be obtained by using the Burg method to compute the reflection coefficients [9]. The Burg method guarantees a minimum phase solution. In the Burg method, the error criterion to be minimized is the sum of the squares of the forward and backward residuals [$f_i(n)$ and $b_i(n)$] at a particular lattice stage. This error term is minimized stage by stage. For a combination filter, the first N_f stages have nonzero reflection coefficients. These are followed by stages with zero-valued reflection coefficients until the Mth stage. The next N_p stages correspond to the pitch part of the filter. Note that as in the cascade connection, the determination of the reflection coefficients in the lattice form proceeds sequentially—the derivation of the coefficients of the formant stages does not take into account the later pitch stages. However, the pitch coefficients do take the preceding formant stages into account.

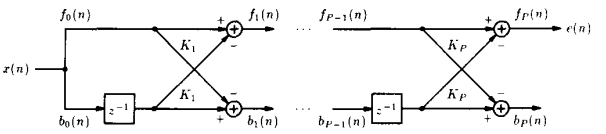

Fig. 3. All-zero lattice filter.

Care must be exercised in a lattice implementation to avoid problems when the pitch lag changes. As the pitch lag changes, the pitch stages change location. For some of the stages, the reflection coefficients change suddenly from a nonzero value to a zero value and vice versa. If the pitch lag increases from one frame to another, the backward residual in the lattice (the signals $b_i(n)$ in Fig. 3) entering the pitch stages will have been filtered by the old pitch coefficients. This can have a detrimental effect on the performance for filters with pitch taps.[1] The remedy is to reset, at each frame boundary, the portion of the backward residual after the formant part of the combination lattice filter to delayed versions of the formant predicted backward residual. This is similar to the remedy used for a lattice pitch filter in [5].

B. Jointly Optimized Cascade Connection

Previous implementations using pitch filters have employed a sequential optimization of the formant and pitch predictors in a cascade connection. Specifically, the formant filter coefficients have been chosen to minimize the intermediate residual signal. This intermediate residual will still have pitch pulses present. The mean-square criterion penalizes the solution for the presence of these relatively high amplitude pulses in the residual. A jointly optimized solution minimizes the mean-square error in the final residual which results after both formant and pitch filtering. This formulation allows the formant filter to ignore pitch pulses in the intermediate residual which will be subsequently removed by the pitch filter.

1) One-Shot Combined Optimization: Under some circumstances, a *one-shot combined optimization* approach to developing an optimal F-P cascade is possible. The one-shot approach involves solving a linear system of equations and requires no iterations. Consider a frame of samples from $n = 0$ to $n = N - 1$. The inputs are the original speech samples $s(n - 1), \cdots, s(n - N_f)$ and samples of the formant predicted residual $d(n - M), \cdots, d(n - M - N_p + 1)$. The analysis essentially decouples the pitch filter from the output of the formant filter as depicted

[1] For formant filters, the same stages are involved in the filtering for each analysis frame. Also, the changes in the reflection coefficients will tend to be less abrupt than those for the pitch stages.

in Fig. 4. To ensure that samples of the formant predicted residual are available beforehand for the analysis, they must depend only on coefficients from previous frames. This imposes the constraint that the pitch lag M be at least as large as the frame length N. The need to have the value of M at least as large as the frame length is a major limitation to the usefulness of this technique. In addition, the requirement that the intermediate residual signal be available beforehand means that the one-shot combined optimization is not possible for a P-F cascade.

For the one-shot combined optimization, the windowed mean-square error is minimized, resulting in a system of equations $\Phi \boldsymbol{c} = \boldsymbol{\alpha}$ where

$$\begin{aligned} \boldsymbol{c}^T &= [a_1, \cdots, a_{N_f}, \beta_1, \cdots, \beta_{N_p}] \\ \Phi &= \sum_{n=0}^{N-1} \boldsymbol{u}^{(n)} \boldsymbol{u}^{(n)T} \\ \boldsymbol{\alpha} &= \sum_{n=0}^{N-1} s(n) \boldsymbol{u}^{(n)} \end{aligned} \tag{5}$$

and

$$\begin{aligned} \boldsymbol{u}^{(n)T} = [&s(n-1), \cdots, s(n-N_f), d(n-M), \\ &\cdots, d(n-M-N_p+1)]. \end{aligned} \tag{6}$$

The system of equations can be written in partitioned form:

$$\left[\begin{array}{ccc|ccc} \phi_{ss}(1,1) & \cdots & \phi_{ss}(1,N_f) & \phi_{sd}(1,M) & \cdots & \phi_{sd}(1,M+N_p-1) \\ \vdots & & \vdots & \vdots & & \vdots \\ \phi_{ss}(N_f,1) & \cdots & \phi_{ss}(N_f,N_f) & \phi_{sd}(N_f,M) & \cdots & \phi_{sd}(N_f,M+N_p-1) \\ \hline \phi_{ds}(M,1) & \cdots & \phi_{ds}(M,N_f) & \phi_{dd}(M,M) & \cdots & \phi_{dd}(M,M+N_p-1) \\ \vdots & & \vdots & \vdots & & \vdots \\ \phi_{ds}(M+N_p-1,1) & \cdots & \phi_{ds}(M+N_p-1,N_f) & \phi_{dd}(M+N_p-1,M) & \cdots & \phi_{dd}(M+N_p-1,M+N_p-1) \end{array}\right] \times \left[\begin{array}{c} a_1 \\ \vdots \\ a_{N_f} \\ \hline \beta_1 \\ \vdots \\ \beta_{N_p} \end{array}\right] = \left[\begin{array}{c} \phi_{ss}(0,1) \\ \vdots \\ \phi_{ss}(0,N_f) \\ \hline \phi_{sd}(0,M) \\ \vdots \\ \phi_{sd}(0,M+N_p-1) \end{array}\right]. \tag{7}$$

where the cross-correlation terms are given by

$$\phi_{xy}(i,j) = \sum_{n=0}^{N-1} x(n-i)y(n-j). \tag{8}$$

The upper left-hand corner of Φ (N_f by N_f) has the terms corresponding to a formant filter acting on the input signal. The lower right-hand corner (N_p by N_p) has the terms corresponding to a pitch filter acting on the formant predicted residual. The other two corners of the matrix contain interaction terms which allow for the combined optimization.

2) Iterated Combined Optimization: The limitation that $M \geq N$ can be circumvented by using an iterative approach to calculate the combined solution. The *iterated combined optimization* method 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 in order to produce the intermediate residual signal. The next iteration uses a combined optimization, but based on the intermediate residual 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. If the error increases during a particular iteration, the process is stopped and the previous solution is kept. Hence, this approach guarantees a mean-square error that is never worse than for the conventional sequential solution.

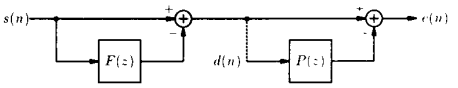

Fig. 4. F-P cascade configuration.

3) Iterative Sequential Optimization: An alternate iterative scheme can be built around a sequential optimization of an F-P cascade. First assume that the formant predictor appears alone or, equivalently, that the pitch predictor coefficients are initially zero. 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 in a conventional F-P cascade. In the *iterated sequential optimization* scheme, the formant filter is reoptimized given the previously determined pitch predictor coefficients. The equations to reoptimize the formant filter tak-

ing into account the pitch filter can be viewed as minimizing a weighted mean-square error criterion. The analysis model is shown in Fig. 5. For the case at hand, the weighting filter $H(z)$ is the pitch prediction error filter.

The analysis model shows an additional input signal $e_o(n)$ which captures the initial conditions and allows for the use of zero initial conditions for all filters. In our case, $e_o(n)$ is the output of the pitch filter due to inputs from previous analysis frames. There is an additional window applied to the intermediate residual signal. For a covariance analysis, the residual window $w_r(n)$ is normally the same as the error window $w_e(n)$. Minimization of the weighted error energy leads to a set of linear equations $\mathbf{\Phi c} = \boldsymbol{\alpha}$, where in this case $\boldsymbol{c}$ is the vector of formant predictor coefficients. The elements $\mathbf{\Phi}$ are given by

$$\phi(i,j) = \sum_{n=-\infty}^{\infty} w_e^2(n)s^{(i)}(n)s^{(j)}(n) \quad \text{for } 1 \le i,j \le N_f, \tag{9}$$

where

$$s^{(i)}(n) = \sum_{m=-\infty}^{\infty} w_r(m)h(n-m)s_w(m-i). \tag{10}$$

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

$$\alpha_i = \sum_{n=-\infty}^{\infty} w_e^2(n)\left[s^{(0)}(n) + e_o(n)\right]s^{(i)}(n). \tag{11}$$

One can note that the effects of the initial conditions due to $e_o(n)$ on the vector $\boldsymbol{\alpha}$ are such that even an autocorrelation solution does not guarantee a stable formant synthesis filter. In the case being studied, the $e_o(n)$ term is due to the relatively long delayed outputs of the pitch filter. It is not appropriate to neglect these for normal analysis frame sizes.

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,

$$s^{(i)}(n) = \sum_{m=0}^{n} h(n-m)s(m-i), \qquad 0 \le n < N. \tag{12}$$

With the weighted error formulation, the formant filter can take into account the effect of the pitch filter. The iterated sequential optimization method first finds the formant filter which minimizes the error taking into account the pitch filter. Next, the formant predicted residual is formed. The pitch filter is optimized based on this residual. These two steps are then iterated. At each step, the overall residual energy decreases and eventually approaches a local minimum.

Some points on the application of the iterated sequential optimization procedure need elaboration. First, the ordering of the predictors is important. In general, different

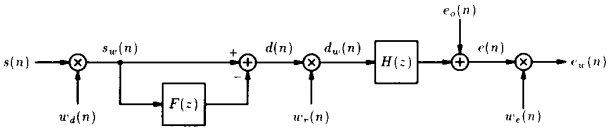

Fig. 5. Analysis model for a prediction filter with error weighting.

solutions will be reached for the F-P and P-F cascades.[2] Second, the initial conditions (starting pitch filter coefficients) affect the solution. We start by solving for a formant filter given a pitch filter with zero valued coefficients. This guarantees results that are at least as good as a conventional sequential solution for an F-P cascade. Third, the procedure is only guaranteed to converge to a local minimum if the optimality criterion is the same for determining both the formant and pitch predictor coefficients.[3] Fourth, the value of M used for the pitch filter is assumed to be constant during the iteration process. If this parameter is changed during the iteration, the convergence properties may be compromised.

IV. Filter Performance

This section presents the experimental results obtained for the combination filter and for F-P cascade connections in which the coefficients are computed by the various joint solutions. The open-loop predictor configuration as depicted in Fig. 1(a) is used to implement the different algorithms. For the purposes of comparison, covariance analyses are 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 and provides an indication of the relative performances of the different approaches by revealing their maximum attainable prediction gains. Later, a practical method of calculating M is combined with the minimum phase constraint for the iterated sequential formulation.

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 decibels 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 decibels 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.

A. Short Frames

Consider first the one-shot combined optimization solution given by (7). In this case, the value of the pitch lag

[2]In the experiments, only the F-P cascade is used due to its superiority over its P-F counterpart in a conventional sequential optimization.

[3]This precludes the use of, say, an autocorrelation analysis for the formant filter and a covariance analysis for the pitch filter.

M must be at least as large as the frame length. A short frame length of 40 samples is used. The minimum and maximum values of M are set to 40 and 140. This combination of parameters will allow us to evaluate the degree of suboptimality of the iterated approaches.

For the iterative sequential scheme, a prediction gain threshold of 0.02 dB or a maximum of 10 iterations was used to terminate the optimization cycle. The resultant F-P cascade formed from the one-shot combined and iterated sequential solutions must outperform the conventional sequentially optimized form for the same analysis conditions. The results are shown in Table I. The combined and iteratively optimized configurations have an improved overall prediction gain. Note that the formant gain drops. However, this drop in formant gain is more than compensated by the increase in pitch gain.

The iterated sequential scheme approaches the (optimal) one-shot combined solution in overall prediction gain. However, one can note that this solution tends to have a higher formant gain and a lower pitch gain than the one-shot combined solution. The iterated sequential method seems to find solutions which have essentially the same overall prediction gain, but a different apportioning of the formant and pitch gains than the one-shot combined solution. The iterative sequential approach outperforms the noniterative sequential solution, again showing a tradeoff toward a higher pitch gain and a lower formant gain.

We note that in all three methods considered in the table, there is a stability problem for the formant filter as well as for the pitch filter due to the fact that a covariance solution has been used. Comments on the relative stability of the corresponding synthesis filters are given in a later section.

B. Longer Frame Sizes

Now, consider a larger frame size of 80 samples. Restricting the pitch lag to be above 80 is not appropriate since the pitch period for voiced speech is frequently below 80. This is especially true for female speakers whose pitch period is usually smaller than for male speakers. For these experiments, the pitch lag assumes values between 20 and 120. 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.

Table II shows the results for 80 sample frames. The iterated sequential method gives larger prediction gains than the iterated combined optimization method at the expense of a slightly larger average number of iterations (5 per frame). For the iterated combined optimization scheme, the iterations are often terminated by a decrease in overall prediction gain (nonmonotonicity of decrease in error).

Combination filters designated as F + P in the table are implemented as both transversal and lattice structures. The transversal form, which takes into account the interaction between the formant part and the pitch part, performs better than the conventional F-P sequential but is inferior to the iterated combined and iterated sequential algorithms. The use of the transversal combination filter is deprecated since the filter frequently becomes nonminimum phase (more than 70 percent of the frames). The table does not show figures for the formant and pitch gains separately for the combination filters. In the transversal form, the prediction gains of the separated components $1 - F(z)$ and $1 - P(z)$ tend to be relatively small, and when added fall far short of the overall prediction gain. The lattice form of a combination filter performs more poorly than the transversal form. This can be attributed to the essentially sequential determination of the reflection coefficients and to the error criterion (sum of forward and backward errors) used in the Burg formulation. It can be said in favor of the lattice implementation that the lattice filter does offer an improvement due to pitch prediction and that no stability problems are encountered. However, the lattice combination filter does not realize the full potential for removing the far-sample redundancies.

TABLE I
PREDICTION GAINS FOR SEQUENTIALLY AND JOINTLY OPTIMIZED PREDICTORS (40 SAMPLE FRAMES). THREE NUMBERS IN AN ENTRY REFER TO 1, 2, AND 3 TAP PITCH FILTERS. A COVARIANCE FORMULATION IS USED FOR ALL CASES

method	formant gain dB	pitch gain dB	overall gain dB
F-P sequential	16.7	4.7 5.9 6.7	21.4 22.6 23.4
F-P iterated sequential	15.0 14.7 14.9	8.5 10.7 10.9	23.5 25.4 25.8
F-P one-shot combined	14.9 13.9 13.7	8.6 11.5 12.3	23.5 25.4 26.0

TABLE II
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

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 transversal	–	–	21.1 21.8 22.6
F+P lattice	–	–	17.0 17.4 17.7
F-P iterated combined	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

The results show that a joint solution can achieve an overall prediction gain that is about 1.5 dB better than the conventional sequential approach. The joint F-P cascade approaches give significantly different solutions than the conventional sequential method as manifested in the lowered formant gain and significantly increased pitch gain. We postpone the investigation of the effects of these differences to a later section in which the stability problems are rectified.

V. STABILITY ISSUES

The synthesis filter used for a combination implementation in transversal form is $H_{F+P}(z) = 1/(1 - F(z) - P(z))$. The denominator polynomial is of high order and has many nonzero coefficients. Experiments show that the percentage of frames with unstable filters is 70, 72, and 75 percent for filters with 1, 2, or 3 pitch coefficients, respectively. The stability was checked by comparing the

TABLE III
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

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

magnitudes of the equivalent set of reflection coefficients with unity. Given the high incidence of instability resulting from the covariance solution and the fact that practical methods to stabilize such a filter are not known, the combination filter is not suitable for applications requiring resynthesis of the speech signal.

Iterated combined and iterated sequential optimization of an F-P cascade leads to more cases of instability than the conventional sequential solution. Table III gives figures for the percentage of frames with unstable synthesis filters for sequential optimization, iterated sequential optimization and iterated combined optimization (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.

The question that is addressed in the next section is whether the increased prediction gain of the joint solutions comes only at the expense of unstable filters.

VI. MINIMUM PHASE JOINT SOLUTION

In this section, an optimization algorithm for an F-P cascade with both a minimum phase formant filter and a minimum phase pitch filter will be described. This algorithm is based on the iterated sequential technique. The formant filter is determined by taking the pitch filter into account as long as its minimum phase property is not sacrificed. In addition, a practical method to determine the pitch lag is used.

The formant and pitch filters are determined as follows. 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 Appendix A. This method gives the optimal pitch lag if the near-sample redundancies have been removed by the formant filter. Practical results in a sequential solution for an F-P cascade show a negligible degradation of the prediction gain when compared to an exhaustive search [5]. Moreover, the immense computation associated with performing all the iterations for every pitch lag and selecting the pitch lag that maximizes the prediction gain (exhaustive search) is avoided. 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].

After the first iteration, the pitch lag is fixed at its initially determined value. 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. An intermediate residual is generated. The pitch filter is again determined based on the updated intermediate residual and stabilized if necessary. These iterations continue until the overall gain increases by less than 0.02 dB.

The formulation given above is based on a common analysis frame for the formant and pitch filter optimization. However, this formulation can be generalized to allow for different formant and pitch frame sizes. Consider a formant analysis frame which is divided into subframes for pitch analysis. In each subframe, the pitch lag and pitch coefficients are optimized. The modification to the iterative procedure involves replacing $h(n)$ in (12) by a time-varying filter. Having different frame sizes for the two parts of the analysis is appropriate in a speech coding context. Good prediction gain is maintained when the formant filter is updated at a slower rate than the pitch filter. Reducing the update rate for the formant filter decreases the side information rate needed to send the predictor coefficients to the decoder.

A. Results for a Minimum Phase Joint Optimization

The resulting prediction gains for a minimum phase iterated sequential solution are shown in Table IV for 80 sample frames. Also included in the table are the prediction gains if the process is stopped at the first iteration (a pure sequential approach with minimum phase filters). 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 II to reveal a drop in overall prediction gain due to the use of stabilized filters and a practical method to find the pitch lag. However, the imposition of the above two constraints does not substantially diminish the prediction gain. The iterated sequential approach continues to outperform the conventional sequential method. Moreover, the gain achieved by the constrained iterative approach still remains above that achieved by the unconstrained F-P sequential algorithm which can have both unstable formant and unstable pitch filters (see Table II).

Fig. 6 shows plots of the prediction gains (3 tap pitch filter) for an utterance ("Thieves who rob friends deserve jail") spoken by a female. The prediction gains are calculated for 80 sample frames. The plots show that the largest increases in prediction gain attained by the iterated approach tend to be in the energetic voiced regions.

Fig. 7 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

TABLE IV
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

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

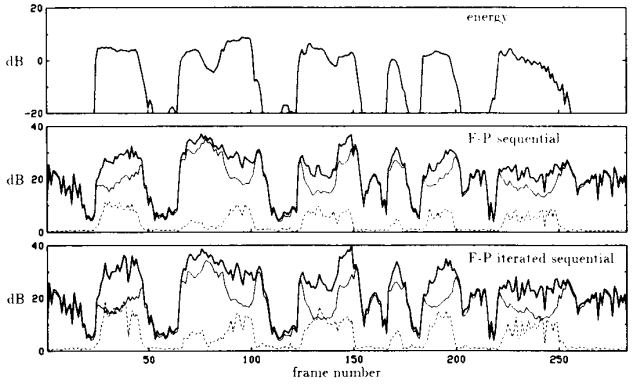

Fig. 6. Prediction gains (3 tap pitch filter), frame by frame. In each of the two lower plots, the upper trace is the overall prediction gain, the middle light trace is the formant prediction gain, and the lower dashed trace is the pitch prediction gain.

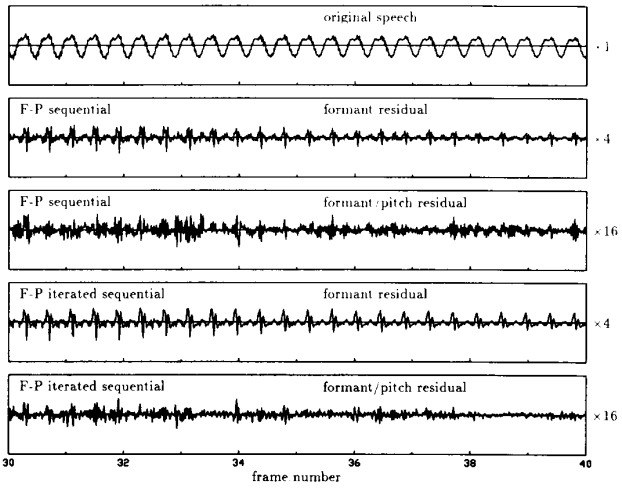

Fig. 7. Prediction residuals (3 tap pitch filter) for a section of voiced speech. Magnification factors are noted beside each plot.

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 forms a final residual having a smaller magnitude and containing fewer spurious bursts than for 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 rudimentary version of a CELP coder. The analysis

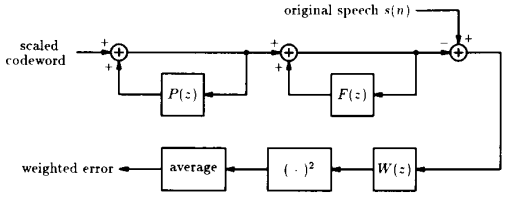

Fig. 8. Calculation of the weighted error in a CELP coder.

frame size remains at 80 samples. A 10th-order formant predictor and a 3 tap pitch predictor were used. Forty sample blocks of the residual were compared to a dictionary of 1024 waveforms consisting of Gaussian random numbers with unit variance. This comparison was performed by the system shown in Fig. 8. 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 [4]. The noise weighting factor α equals 0.8.

Listening 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. 7 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.

Measurements of the signal-to-noise ratio (SNR) that arises when the original speech is compared to the decoded speech were made. For the sentence "Thieves who rob friends deserve jail," the SNR improves by 2.5 dB when the iterated sequential method is used.

VII. SUMMARY AND CONCLUSIONS

This paper has introduced formulations which allow for the joint optimization of the formant and pitch predictors. The final residual can be directly minimized either by a combination implementation or by a cascade of jointly optimized filters. A combination implementation is viewed as a single prediction error filter that removes both near-sample and distant-sample correlations. Either a transversal or lattice structure can be used. The lattice structure is inferior to the transversal form in terms of prediction gain. Although the combination transversal filter outperforms the conventional F-P sequential solution, it suffers from serious stability problems attributable to the covariance method of solution.

For a cascade configuration, a significant advantage in terms of predictor gain exists when the final residual is

directly minimized. The resulting increase in prediction gain comes about by a lowered formant gain that is compensated by a significantly higher pitch gain. However, the limitations of this formulation allow for a one-shot combined solution only for short frame lengths. Iterated schemes can be used to get around the frame length constraints. However, since this approach uses a covariance analysis, it admits both unstable formant pitch synthesis filters.

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 nonminimum phase filters and by stabilizing the pitch filter. The advantages of the iterated sequential method over the one-shot F-P sequential method lie not only in achieving a higher prediction gain but also in producing an overall residual that has diminished pitch spikes. Moreover, experiments with a simple CELP coder show that the resynthesized speech is of better perceptual quality.

Appendix A
Pitch Lag Computation

A practical method to determine the pitch lag has been developed in [5] and is briefly described here. The covariance method applied to a pitch filter leads to a linear system of equations ($\Phi \mathbf{c} = \boldsymbol{\alpha}$) which when written in expanded form is

$$\begin{bmatrix} \phi(M, M) & \cdots & \phi(M, M+N_p-1) \\ \vdots & & \vdots \\ \phi(M+N_p-1, M+N_p-1) & \cdots & \phi(M+N_p-1, M+N_p-1) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{N_p} \end{bmatrix} = \begin{bmatrix} \phi(0, M) \\ \vdots \\ \phi(0, M+N_p-1) \end{bmatrix}, \quad \text{(A.1)}$$

where $\phi(i, j)$ is the correlation for the input to the pitch filter

$$\phi(i, j) = \sum_{n=0}^{N-1} d(n-i)\, d(n-j). \quad \text{(A.2)}$$

Then, $\mathbf{c} = \Phi^{-1}\boldsymbol{\alpha}$ and the resulting mean-squared error is $\epsilon^2 = \phi(0, 0) - \mathbf{c}^T\boldsymbol{\alpha}$. The value of the pitch lag M should be chosen so as to maximize $\mathbf{c}^T\boldsymbol{\alpha}$.

In the case of a one tap pitch filter, $\beta_1 = \phi(0, M)/\phi(M, M)$ and the resulting mean-square error ϵ^2 is

$$\epsilon^2 = \phi(0, 0) - \frac{\phi^2(0, M)}{\phi(M, M)}. \quad \text{(A.3)}$$

The pitch lag M should be chosen so as to maximize $\phi^2(0, M)/\phi(M, M)$. The situation for multitap pitch filters is more complex but is simplified on the assumption that the near-sample based redundancies are essentially removed when formant prediction is performed before pitch prediction (F-P cascade). Then, the off-diagonal terms in the matrix Φ [as in (A.1)] are small and can be neglected. This leads to a diagonal matrix approximation of the system given in (A.1) with $\mathbf{c}^T\boldsymbol{\alpha}$ becoming

$$\mathbf{c}^T\boldsymbol{\alpha} \approx \sum_{m=M}^{M+N_p-1} \frac{\phi^2(0, m)}{\phi(m, m)}. \quad \text{(A.4)}$$

The value of M that maximizes this quantity is chosen as the pitch lag.

References

[1] 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.

[2] J. Makhoul and M. Berouti, "Adaptive noise spectral shaping and entropy coding of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 63–73, Feb. 1979.

[3] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[4] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tampa, FL, Mar. 1985, pp. 25.1.1–25.1.4.

[5] R. P. Ramachandran and P. Kabal, "Pitch prediction filters in speech coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 467–478, Apr. 1989.

[6] ——, "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] 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.

[8] B. W. Dickinson, "Autoregressive estimation using residual energy ratios," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 503–506, July 1978.

[9] J. Makhoul, "Stable and efficient lattice methods for linear prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 423–428, Oct. 1977.

Peter Kabal (S'70–M'75), for a photograph and biography, see p. 478 of the April 1989 issue of this TRANSACTIONS.

Ravi P. Ramachandran, for a photograph and biography, see p. 478 of the April 1989 issue of this TRANSACTIONS.