Codeword Selection for CELP Coders

J.-L..Moncet and P. Kabal

INRS-Télécommunications
3 Place du Commerce
Ile des Soeurs, Que.
CANADA H3E 1H6

July 1987

Rapport technique de l'INRS-Télécommunications no. 87-35

Codeword Selection for CELP Coders

Abstract

This report describes the algorithm used for selecting an excitation waveform for a CELP coder operating at 5 kb/s. Each candidate waveform is used to synthesize a segment of speech. A frequency weighted error criterion is used to find the waveform which regenerates the best output speech. The synthesis operation uses both a pitch synthesis filter and a formant synthesis filter. The pitch synthesis filter is optimized to give the best output speech. This optimization offers a significant improvement over a procedure which uses a pitch filter chosen by analyzing the input speech. Simplified sequential versions of this strategy also give good quality speech. The quantization of the parameters is also considered.

Codeword Selection for CELP Coders

1. Introduction

In Code Excited Linear Prediction (CELP) Coding for speech, a waveform selected from a dictionary of waveforms is used to excite a synthesizer. The output of the synthesizer is the reconstructed speech signal. This report describes the procedure used to select the waveform from the dictionary that produces the "best" match to the original speech signal.

The design philosophy of the overall CELP coder is described in a companion technical report [1]. The present report will concentrate on the waveform selection process itself.

In low bit rate coding, the input speech is processed by two linear predictors to form a residual signal. This is the analysis stage. The first predictor removes near sample redundancies, while the second predictor removes far sample redundancies. The near sample redundancies can be attributed to the vocal tract shaping introduced in speech production. The far sample redundancies can be attributed to the quasi-periodic excitation of the vocal tract in voiced speech. The resulting residual after both stages of prediction is very noiselike and with appropriate gain normalization has a distribution which is nearly Gaussian. It is the noiselike nature of the residual after prediction that motivates the use of a dictionary of so-called stochastic waveforms in CELP coding.

At the synthesis stage, the residual or a coded version of it. is used to excite a pitch synthesis filter and a formant synthesis filter. Conventionally. these synthesis filters are the inverses of the corresponding analysis prediction error filters. A standard coding approach is to quantize the residual signal and use the coded version to excite the synthesis filters. However at low bit rates, sample-by-sample quantization is not possible, since the bit allocation falls below 1 bit per sample. In CELP, a repertoire of excitation signals is used. The excitation waveform that produces the "best" quality speech is found. In a sense, the difference between the residual signal and the chosen excitation signal is the quantization error. The excitation signal is chosen block by block. Viewed as a quantization operation, the quantization is that of a vector of samples. In addition, the waveform selection criterion is based on a frequency weighted difference between the synthesized signal and the speech to be reproduced. The frequency weighting takes into account the effects of masking of the coding noise by the speech formants. As a result, the vector quantizer uses a time-varying error criterion.

As explained so far, the synthesizer filters are derived as part of an analysis operation on the input speech. This is in fact the approach used by early descriptions of CELP coding. In the present work, we consider a refinement in which the pitch synthesis filter will be optimized for the actual excitation waveform selected.

2. CELP Synthesis Stage

The synthesis stage for CELP is shown in Fig. 1. The excitation waveform for the current block (subframe) is $x^{(i)}(n)$. This is scaled by the gain factor G and used to drive the pitch synthesis filter. This filter insert pitch periodicities into the waveform. The output of this filter then drives the formant synthesis filter $H(z)$. This filter shapes the spectral envelope. The resonances of this filter correspond to the formant structure of the speech. Both the pitch synthesis filter and the formant synthesis filter use a transversal filter in a feedback configuration. The coefficients of these filters are updated periodically. The update intervals will be expressed in terms of analysis frames, which are further subdivided into subframes.

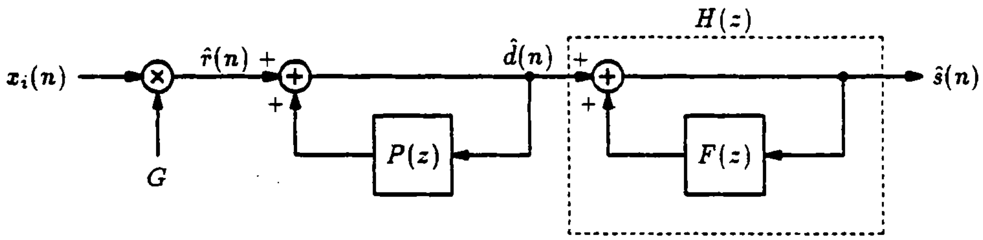

Fig. 1 Speech synthesis model

For the purposes of this study, the formant synthesis filter parameters are derived from the analysis stage. This filter is specified by N_f filter coefficients which are updated once per frame. The pitch synthesis filter can also be derived from the analysis stage. However as part of this study, we also examine pitch synthesis filters which are optimized for the excitation waveform. The pitch synthesis filter is specified by a pitch lag M and a set of N_p filter coefficients. The excitation waveform, the gain factor, and the pitch filter parameters will be updated at the subframe level.

The CELP coder selects the appropriate excitation parameters to be used by the synthesizer. It does this using an analysis-by-synthesis approach; the parameters are selected so as to produce good quality speech when applied to the synthesizer.

3. CELP Analysis-by-Synthesis

Analysis-by-synthesis is used to select the parameters that will be used by the synthesis stage to form an output signal. The criterion used to measure the error in the synthesized signal is based on a frequency weighting.

3.1 Frequency weighted error criterion

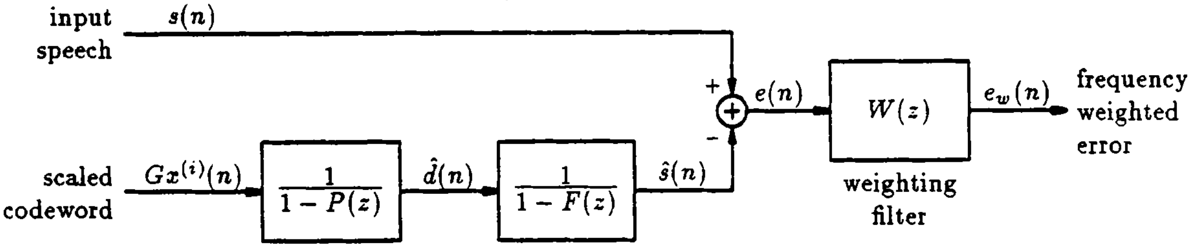

Fig. 2 Frequency weighted error criterion

In order to optimize the synthesized waveform, a frequency weighted error criterion is used. The model used is shown in Fig. 2. The lower part of the figure synthesizes a speech segment. An error signal is formed as the difference between the input speech and the synthesized speech. This is then passed through a frequency weighting filter. The final error measure is the mean-square value of the weighted error. The set of synthesis parameters is chosen to minimize this value.

The transfer function of the weighting filter is given by

$$\begin{aligned} W(z) &= \frac{1-F(z)}{1-F(\gamma z)} \\ &= \frac{H(\gamma z)}{H(z)} , \end{aligned} \tag{1}$$

where γ is a bandwidth expansion factor. The role of the weighting filter is to concentrate the coding noise in the formant regions where it is effectively masked by the speech signal. By doing so, the noise at other frequencies can be lowered to reduce the overall perceived noise. The value chosen for γ is 1/0.75. Note that the weighting filter is related to the formant synthesis filter, and hence is time-varying.

With the given form of the weighting filter, the calculation of the frequency weighted filter can be rearranged as shown in the block diagram in Fig. 3. In this arrangement, the formant synthesis filter and the weighting filter have been combined to form a bandwidth-expanded synthesis filter. The notation for the signals uses primes (e.g. $s'(n)$) to indicate signals which use the bandwidth-expanded synthesis filter and carets (e.g. $\hat{s}(n)$) to indicate coded signals. Note that this bandwidth-expanded synthesis filter is used only in the process of selecting the optimum synthesis parameters. The decoder will use the normal formant synthesis filter.

3.2 Selecting the synthesis parameters

The parameters that are to be selected consist of the input waveform and the filter parameters. The input waveform is $Gx^{(i)}(n)$. The waveform index i and the gain factor G are chosen to produce the best quality speech. The pitch filter is specified by a pitch lag and the filter coefficients. The

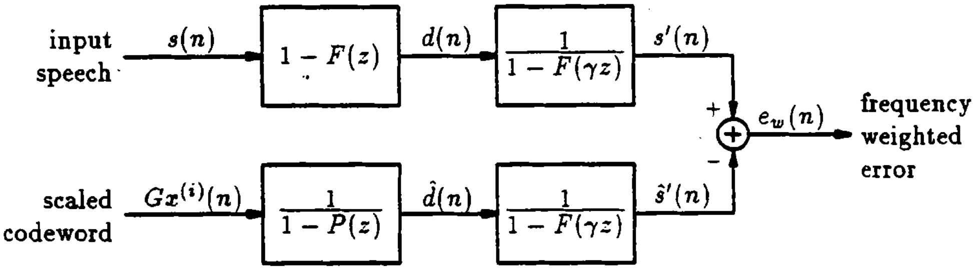

Fig. 3 Alternative model for the calculation of the frequency weighted error

pitch filter parameters are also chosen optimize the synthesized speech. This optimization procedure is described in the next section.

Conventional coders derive the synthesis filter parameters by analyzing the input speech. In this process, the input speech is passed through prediction error filters. These filters are the inverses of the synthesis filters and use the same coefficient values. The coefficients are chosen to minimize the energy of the prediction residual. The formant filter parameters are chosen in this way.

A note on the effect of updating the formant synthesis filter is in order. Normally the bandwidth-expanded synthesis filters in the upper and lower branches of Fig. 3 are updated in synchronism. The filters are updated at the beginning of each frame and have a constant response for the duration of the frame. The filtering action can be expressed in terms of a direct convolution with the filter impulse response. However, due to the all-pole nature of the formant synthesis filter, a recursive formulation for the filtering may be more efficient. The two formulations have the same steady-state output but differ when the filters are time-varying. Differences can be ascribed to the different initial conditions at the frame boundaries. However in the subsequent derivations, the effect of the initial condition will be absorbed into the term which is not affected by the optimization. With zero initial conditions and filter response which is held constant over the subframe, the two filter formulations give identical outputs.

4. Optimization of the Synthesis Parameters

The weighted error can be expressed as

$$\begin{aligned} e_w(n) &= s'(n) - \hat{s}'(n) \\ &= s'(n) - \sum_{k=-\infty}^{\infty} \hat{d}(k)h'(n-k) \ , \end{aligned} \tag{2}$$

where $\{h'(k)\}$ denotes the impulse response of the bandwidth-expanded synthesis filter. As noted earlier, the filter response is actually time-varying. However, the focus here is on the subframe interval during which the filter response is held constant. The term $s'(n)$ will be calculated separately and does not depend on the optimization of the synthesis parameters. Note that the all-pole bandwidth-expanded synthesis filter has a causal impulse response.[†] The bandwidth-expanded filter is derived by analyzing the input speech.

The waveform selection process involves computing the energy of the weighted error

$$\varepsilon = \sum_{n=0}^{N-1} e_w^2(n) \ . \tag{3}$$

The length of the subframe is indicated as N. The optimal selection of the pitch lag and the excitation waveform involves an exhaustive search among all possible pairs (M, i) representing pitch lags and waveform indices. For each such pair, the gain factor G and the coefficients of the pitch filter are chosen to minimize the frequency weighted mean-square error.

It is convenient to rewrite the weighted error $e_w(n)$ in a form with three terms,

$$e_w(n) = s'(n) - \sum_{k=-\infty}^{-1} \hat{d}(k)h'(n-k) - \sum_{k=0}^{\infty} \hat{d}(k)h'(n-k) \ . \tag{4}$$

The second term is the contribution from past codewords. This term is shown in the convolution form. However, if the filter is implemented in recursive form, this term should include the output due to the initial conditions at the subframe boundary. The last term is the contribution from the codeword for the present analysis interval with zero initial conditions for the filter. It is only this term that is optimized with respect to the choice of parameters for the present analysis interval. For convenience, the terms that are not affected by the optimization are lumped into a single term,

$$s''(n) = s'(n) - \sum_{k=-\infty}^{-1} \hat{d}(k)h'(n-k) \ . \tag{5}$$

Now the weighted error can be written as

$$e_w(n) = s''(n) - \sum_{k=0}^{\infty} \hat{d}(k)h'(n-k) \ . \tag{6}$$

The limits of the convolution sum serve to select a portion of the signal. It is useful to define a window function $w_{[N_L,N_U)}(n)$ which selects the interval $[N_L, N_U)$,

$$w_{[N_L,N_U)}(n) = \begin{cases} 1 & N_L \le n < N_U \\ 0 & \text{otherwise} \ . \end{cases} \tag{7}$$

This window function will be used to allow the convolution sum to run from $-\infty$ to $+\infty$,

$$e_w(n) = s''(n) - \sum_{k=-\infty}^{\infty} w_{[0,N)}(k)\,\hat{d}(k)\,h'(n-k) \ , \quad 0 \le n < N \ . \tag{8}$$

The window function selects the portion of the signal $\hat{d}(n)$ in the interval $0 \le n < N$.

[†] Since the filter is causal, the upper limit of the convolution sum could be changed to n.

The output of the pitch synthesis filter can be written as the weighted sum of the waveform from the dictionary and delayed previous outputs,

$$\hat{d}(n) = G x^{(i)}(n) + \sum_{j=1}^{N_p} \beta_j \hat{d}(n - M - j + 1) \ . \tag{9}$$

Substituting this expression into the formula for the weighted error gives

$$\begin{aligned} e_w(n) &= s''(n) - G \sum_{k=-\infty}^{\infty} w_{[0,N)}(k) x^{(i)}(k) h'(n-k) - \sum_{j=1}^{N_p} \beta_j \sum_{k=-\infty}^{\infty} w_{[0,N)}(k) \hat{d}(k-M-j+1) h'(n-k) \\ &= s''(n) - G\, \tilde{x}^{(i)}_{[0,N)}(n) - \sum_{j=1}^{N_p} \beta_j \, \tilde{d}_{[0,N)}(n, M+j-1) \ , \end{aligned} \tag{10}$$

where filtered versions of $x^{(i)}(n)$ and $\hat{d}(n)$ have been defined as

$$\begin{aligned} \tilde{x}^{(i)}_{[N_L,N_U)}(n) &= \sum_{k=-\infty}^{\infty} w_{[N_L,N_U)}(k)\, x^{(i)}(k) h'(n-k) \\ \tilde{d}_{[N_L,N_U)}(n,m) &\doteq \sum_{k=-\infty}^{\infty} w_{[N_L,N_U)}(k)\, \hat{d}(k-m) h'(n-k) \ . \end{aligned} \tag{11}$$

The window applied to the input signal $x^{(i)}(n)$ is superfluous, since the input signal is zero outside the interval $[0, N)$. However it is included to emphasize the time-limited nature of the signal.

4.1 Solution for $M \geq N$

The values of the gain factor G and the coefficients $\{\beta_i\}$ which minimize the squared-error ε are to be found. Appendix A derives the matrix equations for a covariance solution to minimize ε. In matrix form, $\mathbf{\Phi a} = \mathbf{b}$, where the matrix of autocorrelation terms is

$$\mathbf{\Phi} = \sum_{n=0}^{N-1} \mathbf{v}^{(n)} \mathbf{v}^{(n)^T} \ , \tag{12}$$

and

$$\mathbf{v}^{(n)} = \begin{bmatrix} \tilde{x}^{(i)}_{[0,N)}(n) \\ \tilde{d}_{[0,N)}(n, M) \\ \tilde{d}_{[0,N)}(n, M+1) \\ \vdots \\ \tilde{d}_{[0,N)}(n, M+N_p-1) \end{bmatrix} . \tag{13}$$

The coefficient vector $\mathbf{a}$ is defined as

$$\mathbf{a} = \begin{bmatrix} G \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{N_p} \end{bmatrix} . \tag{14}$$

The righthand side vector of cross-correlations is given by

$$\mathbf{b} = \sum_{n=0}^{N-1} s''(n)\mathbf{v}^{(n)} , \qquad (15)$$

Consider the case that $M \geq N$. The filtered signal $\tilde{d}_{[0,N)}(n, m)$ which appears in $\mathbf{v}^{(n)}$ depends only on the signal $\hat{d}(n)$ for $n < 0$. This part of the signal is known from the previous subframe. For a pitch lag larger than the frame size, the matrix $\mathbf{\Phi}$ and the righthand side vector $\mathbf{b}$ are known quantities. The determination of the optimum coefficients involves solving the set of linear simultaneous equations.

5. Performance with the Optimized Synthesis Parameters

The solution method proposed finds the jointly optimal values of the waveform index i, the pitch lag M, the gain G, and the pitch filter coefficients $\{\beta_i\}$. This is accomplished by finding the optimal coefficient vector for each pair (i, M). The pitch lag M is constrained to be at least as large as N.

The joint optimization will be compared with a strategy which chooses the pitch filter parameters (M and $\{\beta_i\}$) by analyzing the input speech. For this comparison the sampling frequency is 8 kHz. The frame size is 80 samples (10 ms) for the formant filter update and the subframe size is 40 samples (5 ms) for the waveform selection, and gain and pitch filter update. The pitch lag takes on values from 40 to 103 (5–12.9 ms). Only a single pitch coefficient will be used. Both the gain and the pitch coefficient are unquantized. The excitation waveform $x^{(i)}(n)$ is chosen from a repertoire of 32 waveforms. These parameters are appropriate for a CELP coder operating near 5 kb/s.

Figure 4(a) shows a segment of an utterance spoken by a male. The lower part of the figure shows the frequency weighted SNR (signal-to-noise ratio) in dB for a CELP coder using pitch synthesis parameters determined by analyzing the input speech (thin line) and using synthesis parameters optimized for the synthesis stage. The average noise weighted SNR increases from 3.9 dB to 6.6 dB when the optimized parameters are used. Figure 5 compares the spectra of the coded sequences to that of the original speech for a 40 ms interval taken from the utterance. It is seen that the harmonic structure is better reproduced with the optimized coefficients. Also the gross spectral information tends to be preserved even towards the high frequencies. The scheme which uses the filter developed at the analysis stage offers a poor match especially around the zeroes in the spectral envelope.

The optimal selection of the excitation, gain and pitch parameters is a considerable improvement over methods used in previous CELP coders. As more structure is added to the excitation waveform through the selection of an optimal pitch filter, the size of the codebook becomes less critical. With the proposed scheme operating on blocks of 5 ms duration, dictionary sizes as small as 16 or 32

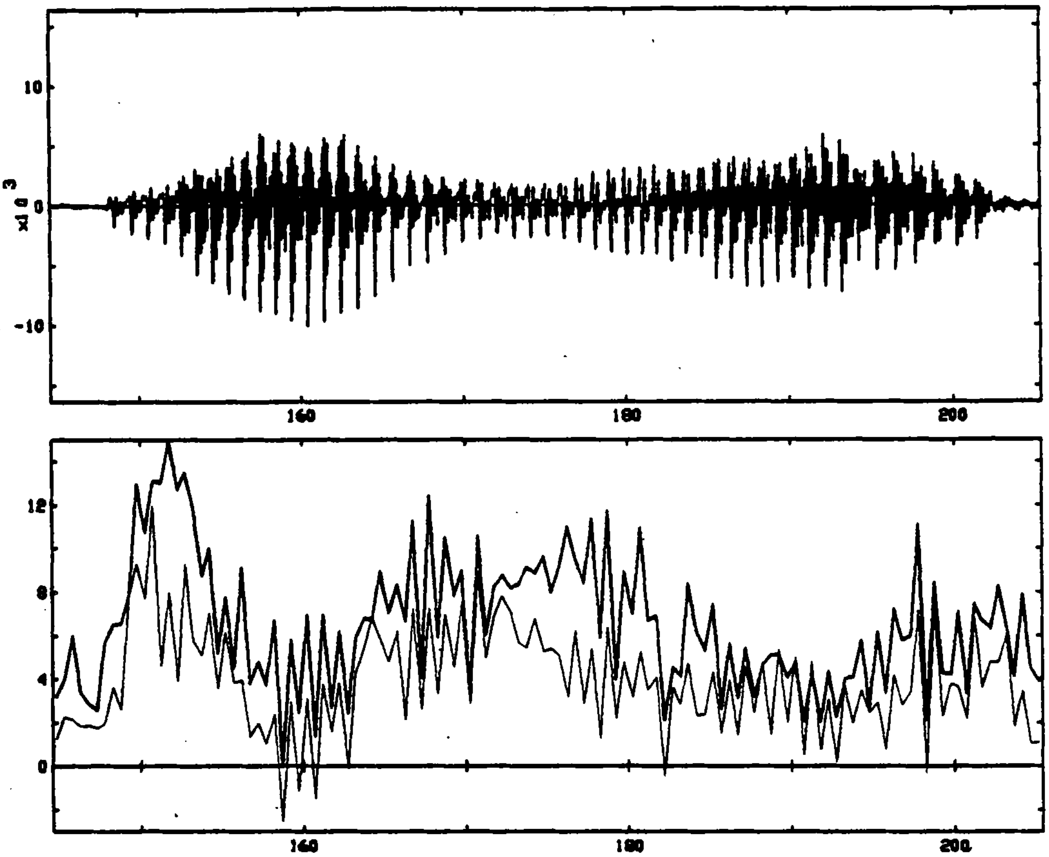

Fig. 4 Comparison of the performance for CELP with pitch filter parameters determined by analyzing the input speech and with parameters optimized for synthesis.
(a) Time waveform.
(b) Segmental noise weighted SNR (thin line for parameters developed by analyzing the input speech, thick line for parameters optimized for the synthesis stage).

waveforms produce reasonable speech quality. The quality is directly comparable to a CELP coder with 1024 waveforms but which uses a pitch filter derived at the analysis stage. The negative aspect of the optimal selection scheme is computational complexity, even for small codebook sizes.

5.1 Sequential approach to the determination of the excitation waveform

In order to reduce the computational load of the parameter optimization procedure, sequential approaches for determining the synthesis parameters are considered. The sequential algorithm determines the pitch lag M using a zero input from the dictionary ($G = 0$). Given that the gain factor G is zero, the equations of Section 4 reduce in complexity. The equations written in matrix form

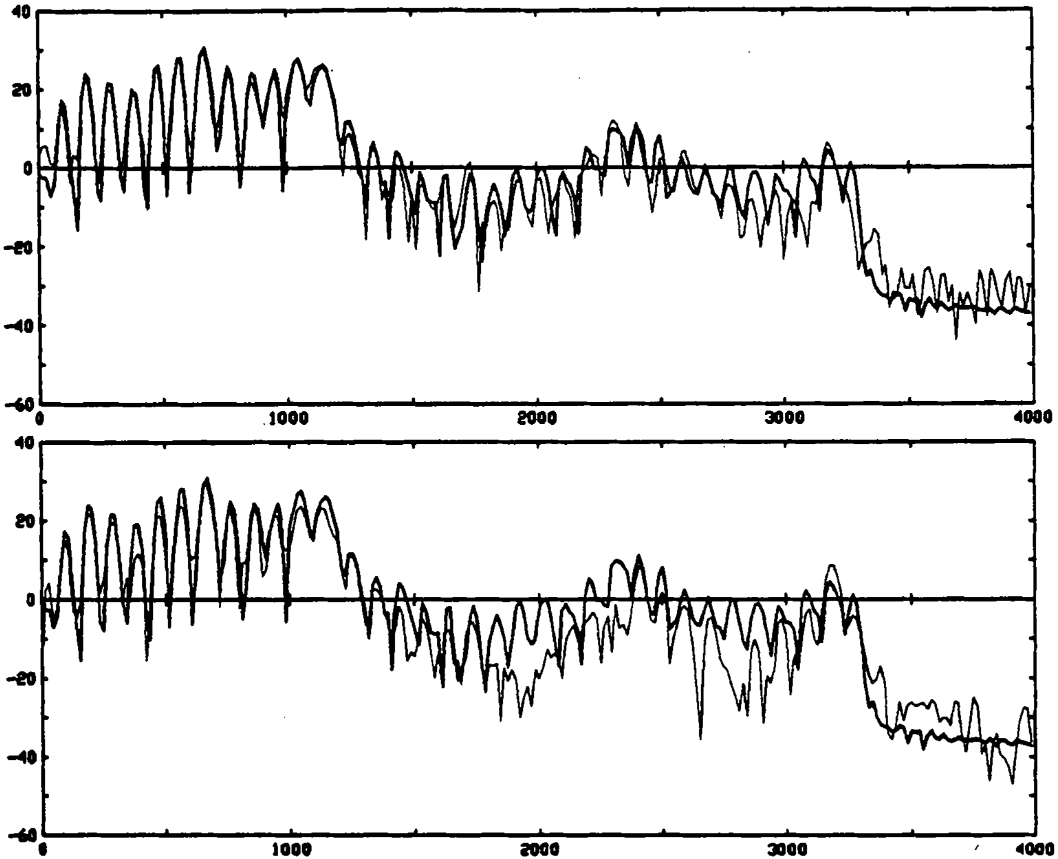

Fig. 5 Comparison of the spectra of the reconstructed signals for a CELP coder (for frames 162–166, c.f. Fig. 4). In both figures, the thick line is the spectrum of the input speech.
(a) parameters optimized for the synthesis stage and
(b) parameters developed at the analysis stage.

are $\mathbf{\Phi}\boldsymbol{\beta} = \mathbf{b}$, where the the vector $\mathbf{v}^{(n)}$ and $\boldsymbol{\beta}$ are defined as

$$\mathbf{v}^{(n)} = \begin{bmatrix} \bar{d}_{[0,N)}(n, M) \\ \bar{d}_{[0,N)}(n, M+1) \\ \vdots \\ \bar{d}_{[0,N)}(n, M+N_p-1) \end{bmatrix} , \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{N_p} \end{bmatrix} . \tag{16}$$

The matrix $\mathbf{\Phi}$ and the vector $\mathbf{b}$ are defined as before. The resulting squared error is (see Appendix A)

$$\varepsilon_{\text{opt}} = \sum_{n=0}^{N-1} s^2(n) - \mathbf{b}^T \mathbf{\Phi}^{-1} \mathbf{b} . \tag{17}$$

The second term in this expression is a function of the pitch lag M through the dependence of $\mathbf{b}$ and $\mathbf{\Phi}$ on $\mathbf{v}^{(n)}$ which in turn depends on M. The optimal value of M can be found by a maximization over the allowable pitch range of the quantity $\mathbf{b}^T \mathbf{\Phi}^{-1} \mathbf{b}$. For a single tap pitch filter, the quantity to be maximized is

$$\max_M(\mathbf{b}^T \mathbf{\Phi}^{-1} \mathbf{b}) = \max_M \left(\frac{\left[\sum_{n=0}^{N-1} s''(n) \bar{d}_{[0,N)}(n, M) \right]^2}{\sum_{n=0}^{N-1} [\bar{d}_{[0,N)}(n, M)]^2} \right) . \tag{18}$$

With the optimum lag determined for a zero excitation, the lag is kept fixed at this value. Two variants of the basic scheme will be presented. In the first variant, the other parameters of the synthesizer are determined by searching over waveform indices. For each waveform index, the optimum gain and pitch coefficients (assuming the pitch lag already determined) are found. In the second variant, the pitch coefficients are also determined for zero excitation. Keeping the pitch filter fixed (both lag and coefficients), the search is conducted over waveform indices. For each index, the optimum gain is found, assuming the other synthesis parameters are fixed.

One interpretation of the operation of the sequential approaches is as follows. The excitation signal which is used to drive the formant synthesis filter is composed of two components. The first is a scaled and delayed version of the previous excitation signal. In voiced speech, this approach supplies the pitch component. The gain scaled waveform from the dictionary fills in details that are missing in the excitation signal. It also supplies the startup component for the pitch excitation in transition regions (unvoiced or silence to voiced).

One can expect that the performance of the sequential approach will degrade somewhat from the optimal joint solution. The first variant can be viewed as the same as the second variant, but with the pitch coefficients reoptimized for each waveform. The methods were compared using unquantized coefficient values. The system parameters are as before.

The first variant produces speech which can be described as smoother than the optimal method, but which lacks a certain fullness. In addition, the energy variations are not rendered quite as accurately. The second variant produces a pitch coefficient which is within 10% of the value given by the first variant in steady voiced speech. Larger differences are observed in silence and transition regions as well as in voiced segments with rapid formant changes. The reproduced speech is much the same for the two variants. There are isolated degradations apparent using the second method which are not present using the first method.

The results point to the fact that the first variant may be the method of choice. The overall differences between the optimal scheme and this method are small enough that the computational savings associated with the sequential approach are attractive.

6. Quantization of the Synthesis Parameters

In this section the problem of coding the synthesizer parameters is considered. The parameters under consideration are the pitch lag, the pitch coefficient, and the gain factor. The variation of these parameters is shown in Fig. 6 for a male utterance. For the purposes of this study the formant synthesis filter parameters are not quantized. Methods to quantize these have been described in [2]. The waveform index is represented with 5 bits (32 waveforms).

The goal is to produce a CELP coder with a bit rate near 4800 b/s. The bit assignment used is shown in Table 1. The total bit rate for coding the parameters is 4000 b/s. This leaves 800 b/s for coding the formant synthesis filter parameters.

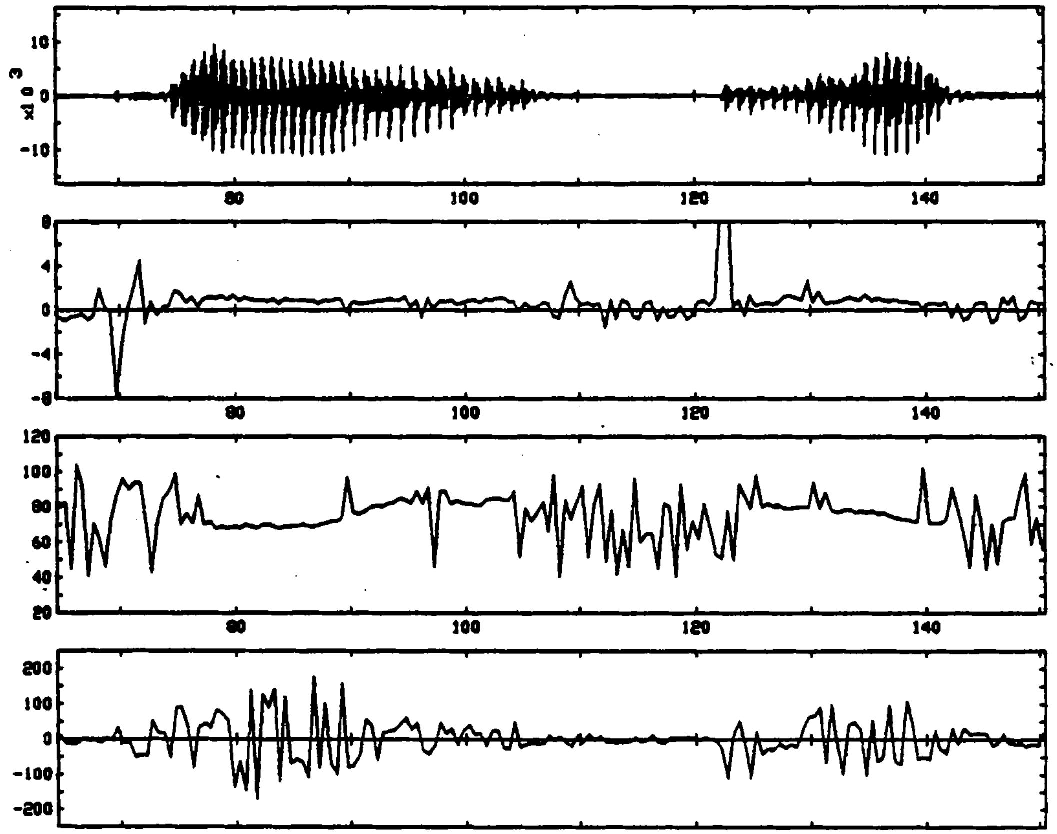

Fig. 6 Parameter tracks.
(a) speech waveform (male utterance),
(b) pitch filter coefficient,
(c) predictor lag, and
(d) the gain factor.

6.1 Pitch lag

The pitch lag ranges from 40 to 103 and is represented with 6 bits. Experiments show that the pitch lag tends to vary smoothly in voiced segments, and only occasionally departs from the smooth trajectory. However, in unvoiced speech the pitch lag tends to jump around. Attempts were made to

Parameter	Transmission Rate	
	bits/subframe	bits/second
M	6	1200
β	4	800
i	5	1000
G	5	1000
total	20	4000

Table 1 Bits allocation for the excitation parameters for each 40 sample sub-frame.

reduce the number of bits needed to code this parameter. One variation allowed the pitch lag to take on only even values (two sample resolution). Another version allowed for single sample resolution in the neighborhood of the previous pitch lag values, but with two sample resolution further away. None of these attempts to reduce the pitch lag resolution performed satisfactorily. The parameter track in Fig. 6(c) shows that the pitch varies widely in silence or unvoiced regions before settling down in the voiced region. The reduced resolution schemes tend to have problems locking in to the correct pitch lag at transitions from silence or unvoiced speech to voiced speech.

6.2 Pitch coefficient

Only a single pitch coefficient is used in the CELP coder. Figure 7 shows the histogram of pitch filter coefficient values. The histogram is obtained as the composite of 10 speech utterances (both male and female speakers) consisting of 4752 subframes. However, the histogram does not give information on the relative importance of different pitch filter values. For instance, the negative coefficient values tend to occur in speech regions with low energy in which the pitch filter does not effect the output speech quality (see Fig. 6(b)).

The pitch coefficient is quantized using 4 bits. Of the 16 levels, only 3 are used to code negative values. The quantizer levels for the positive and negative quantization regions were designed independently. The dynamic range of the quantizer was chosen based on subjective evaluations. The very large pitch coefficient values tend to occur in transition regions (silence to speech) in which the pitch filter does not contribute much to the quality. The levels were determined so as to minimize the mean-squared error of the pitch coefficient.[†] The resulting quantization levels are shown in Table 2.

The quantization of the pitch coefficient interacts with the method to select the pitch coefficient. Recall that two variants of a sequential procedure were used to select the parameters. In the first, the pitch coefficient and gain factor were chosen together. In the second, the pitch coefficient was

[†] A mean-squared error criterion can be justified for maximizing the predictor gain in a pitch predictor which follows a formant predictor [3]. Here we are working on the pitch synthesis filter, using the same criterion to quantize the coefficient.

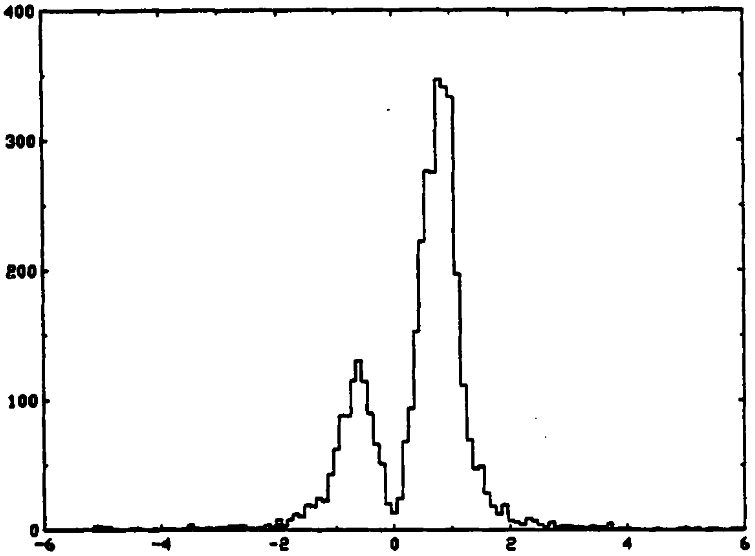

Fig. 7 Histogram of pitch predictor coefficient values.

Decision Levels	Output Levels
−1.02	−1.26
−0.59	−0.79
0.00	−0.40
0.26	0.17
0.41	0.34
0.55	0.49
0.67	0.61
0.78	0.73
0.89	0.83
1.00	0.95
1.13	1.06
1.29	1.20
1.50	1.39
1.77	1.61
2.21	1.93
	2.49

Table 2 Optimal quantizer for pitch predictor coefficient.

chosen first, and then the gain factor determined. The second variant allows for the quantization of the pitch coefficient before the gain factor is chosen. Then the gain factor can compensate for quantization errors. Experiments show that this is indeed so. The speech quality remains close to the reference system which uses the variant one approach with no quantization of the parameters.

6.3 Gain factor

The parameter track for the gain factor shown in Fig. 6(d) shows that the sign of the gain factor jumps around erratically. One can consider that the sign of the gain contributes to the expansion of the waveform dictionary size by a factor of two. With this viewpoint, the gain would be positive and the dictionary of waveforms would consist pairs of opposite sign waveforms.

For coding purposes, the sign and the magnitude of the gain are separated. The magnitude of the gain is coded using a differential approach. The histogram for the difference values is shown in Fig. 8. The histogram was been obtained from the 10 utterance data base used earlier. The magnitude gain difference quantizer uses 16 non-uniformly spaced levels. The quantizer levels are shown Table 3. The quantizer levels were determined so as to minimize the mean-square quantization error for the gain parameter. One might be tempted to use a symmetric quantizer with a zero output level. However, experiments show that for the given dynamic range, coding noise produced by a 15 level quantizer (especially in silence regions) is annoying. For that reason a 16 level quantizer is to be preferred.

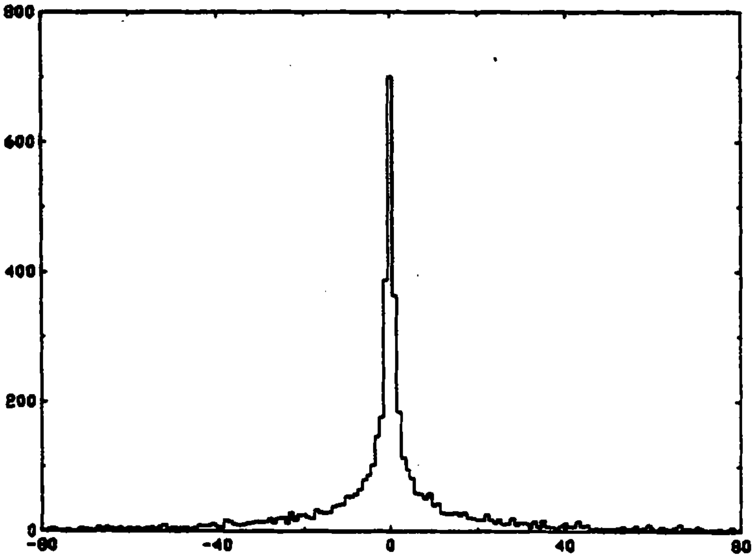

Fig. 8 Histogram of the difference values for the magnitude of the gain.

7. Solution for $M < N$

The parameter coding described in the previous section is effective. Only slight differences in quality can be ascribed to quantization. In fact, the quantization tends to mask some artifacts of the unquantized system. However, the limitation that the pitch lag be greater than the subframe size

Decision Levels	Output Levels
−65.18	−72.28
−51.90	−58.08
−39.88	−45.72
−29.00	−34.05
−19.33	−23.95
−10.93	−14.71
−4.23	−7.15
0.00	−1.30
4.23	1.30
10.93	7.15
19.33	14.71
29.00	23.95
39.88	34.05
51.90	45.72
65.18	58.08
	72.28

Table 3 Optimal quantizer for the magnitude difference gain values.

causes some problems for high pitched female speech. This effect is present in both the quantized and unquantized versions.

The pitch period in our female samples can become as low as 28 samples (3.5 ms, corresponding to a 285 Hz pitch frequency). One can argue that pitch doubling can capture this short pitch period. However, some wavering in the speech can be observed whenever the pitch period hovers around the 40 sample value. This is caused by the pitch lag changing suddenly between its fundamental value and its pitch doubled value. In addition, one can note an imprecision in the harmonic structure when pitch doubled values are used.

One solution to the problem is to reduce the subframe size. However, this has a severe bit rate penalty associated with it. Instead, a modification of the algorithm to find the optimum pitch lag and pitch coefficient has been formulated.

The basic problem in solving for the gain and pitch coefficient for lags less than the subframe size is the nonlinear nature of the resulting equations. The formulation developed earlier holds for $M > N$. However, the equations become nonlinear in the coefficients for $M > N$. This is due to the fact that both the matrix $\mathbf{\Phi}$ and the vector $\mathbf{b}$ contain terms in $\tilde{d}(n)$ for $n \geq 0$. These terms in turn depend on the coefficients. The general solution of the nonlinear set of equations is impractical.

Consider the case that a single pitch coefficient is being sought ($N_p = 1$) for a zero signal from the dictionary ($G = 0$). Also let the pitch lag lie in the interval $N/2 \leq M < N$, where N is the subframe size.

The excitation signal takes on one of two forms

$$\hat{d}(n) = \begin{cases} \beta\hat{d}(n-M) & 0 \le n < M \\ \beta^2\hat{d}(n-2M) & M \le n < N \ . \end{cases} \tag{19}$$

The weighted error is (Eq. 6),

$$e_w(n) = s''(n) - \sum_{k=\infty}^{\infty} \hat{d}(k)h'(n-k) \ . \tag{20}$$

Consider the values of $e_w(n)$ for $0 \le n < M$,

$$\begin{aligned} e_w(n) &= s''(n) - \sum_{k=-\infty}^{\infty} w_{[0,M)}(k)\,\hat{d}(k)\,h'(n-k) \\ &= s''(n) - \beta\tilde{d}_{[0,M)}(n,M) \end{aligned} \qquad 0 \le n < M \ . \tag{21}$$

For $M \le n < N$, the expression involves an extra term,

$$\begin{aligned} e_w(n) &= s''(n) - \sum_{k=0}^{M-1} \hat{d}(k)h'(n-k) - \sum_{k=M}^{\infty} \hat{d}(k)h'(n-k) \\ &= s''(n) - \beta\tilde{d}_{[0,M)}(n,M) - \beta^2\tilde{d}_{[M,N)}(n,2M) \end{aligned} \qquad M \le n < N \ . \tag{22}$$

The squared-error sum can be broken into two parts

$$\begin{aligned} \varepsilon &= \sum_{n=0}^{M-1} e_w^2(n) + \sum_{n=M}^{N-1} e_w^2(n) \\ &= \sum_{n=0}^{M-1} [s''(n) - \beta\tilde{d}_{[0,M)}(n,M)]^2 + \sum_{n=M}^{N-1} [s''(n) - \beta\tilde{d}_{[0,M)}(n,M) - \beta^2\tilde{d}_{[M,N)}(n,2M)]^2 \\ &= \sum_{n=0}^{N-1} [s''(n)]^2 - 2\beta\sum_{n=0}^{N-1} s''(n)\tilde{d}_{[0,M)}(n,M) + \beta^2\sum_{n=0}^{N-1} [\tilde{d}_{[0,M)}(n,M)]^2 \\ &\qquad - 2\beta^2\sum_{n=M}^{N-1} s''(n)\tilde{d}_{[M,N)}(n,2M) + 2\beta^3\sum_{n=M}^{N-1} \tilde{d}_{[0,M)}(n,M)\tilde{d}_{[M,N)}(n,2M) \\ &\qquad + \beta^4\sum_{n=M}^{N-1} [\tilde{d}_{[M,N)}(n,2M)]^2 \ . \end{aligned} \tag{23}$$

This is a nonlinear equation in the single parameter β. In fact, setting the derivative to zero gives a cubic in β which can be solved in closed form. However, the solution of the cubic involves the computation of transcendental functions. Also note that adding the input term with the gain factor G would add greatly to the complexity by giving rise to nonlinear cross terms.

The proposed method for finding the optimum value for β takes a short cut based on using the quantized values for β. In this scheme, the sum terms are precomputed. Each of the possible quantized values for β is substituted into the equation. The value of β which gives the smallest value for ε is chosen. For a relatively small number of quantized values, this approach is computationally more efficient than solving the cubic in β directly.

A second method for calculating the pitch coefficient for $M < N$ is based on more empirical foundations. In this scheme, the past pitch filter output is periodically continued,

$$\hat{d}(n) = \begin{cases} \beta\hat{d}(n-M) & \text{for } 0 \le n < M \ , \\ \beta\hat{d}(n-2M) & \text{for } M \le n < N \ . \end{cases} \tag{24}$$

This scheme embodies an automatic pitch doubling for part of the subframe. With this formulation, the equation for the squared error is a quadratic in β,

$$\begin{aligned} \varepsilon = \sum_{n=0}^{N-1} [s''(n)]^2 - 2\beta \sum_{n=0}^{N-1} s''(n)\bar{d}_{[0,M)}(n,M) + \beta^2 \sum_{n=0}^{N-1} [\bar{d}_{[0,M)}(n,M)]^2 \\ - 2\beta \sum_{n=M}^{N-1} s''(n)\bar{d}_{[M,N)}(n,2M) + 2\beta^2 \sum_{n=M}^{N-1} \bar{d}_{[0,M)}(n,M)\bar{d}_{[M,N)}(n,2M) \\ + \beta^2 \sum_{n=M}^{N-1} [\bar{d}_{[M,N)}(n,2M)]^2 \ . \end{aligned} \tag{25}$$

Setting the derivative of the squared error to zero gives a linear equation in β.

7.1 Results for an expanded pitch lag range

Figure 9 gives an example of the results obtained when the pitch lag is allowed to fall below the subframe size of 40 samples. In this portion of a female utterance, the pitch frequency is steady at about 225 Hz, corresponding to pitch lags of 35 or 36 samples . In the original scheme, lag values smaller than the 40 sample subframe length are not allowed and the pitch predictor is forced to operate around multiples of the fundamental period instead. The consequences are readily seen in Fig. 9(b) where the spectrum of the reconstructed speech displays excessive energy concentration around harmonics of 75 Hz and 112.5 Hz which show up in between the main spectral peaks. Subjectively, this is heard as a lack of homogeneity and smoothness in the reconstructed speech. The solution of the nonlinear equation in β by trial and error was applied to this segment of speech. Fig. 9(c) shows that the scheme without the pitch lag restriction is quite effective in cancelling those spurious peaks. The resulting speech is considerably improved in quality.

The method employing the periodic continuation of the pitch filter output was also tried. From the formulation, one can see that method does not allow for pitch pulses in a subframe which change in amplitude from one pulse to another. It can underestimate the impact of coefficient values greater than one. This represents a potential cause for local degradations of the synthetic speech. Subjective comparisons involving the two methods show that the second gives slightly poorer results. More obvious was the presence of occasional artifacts in the reconstructed speech due to sudden bursts of the high frequencies.

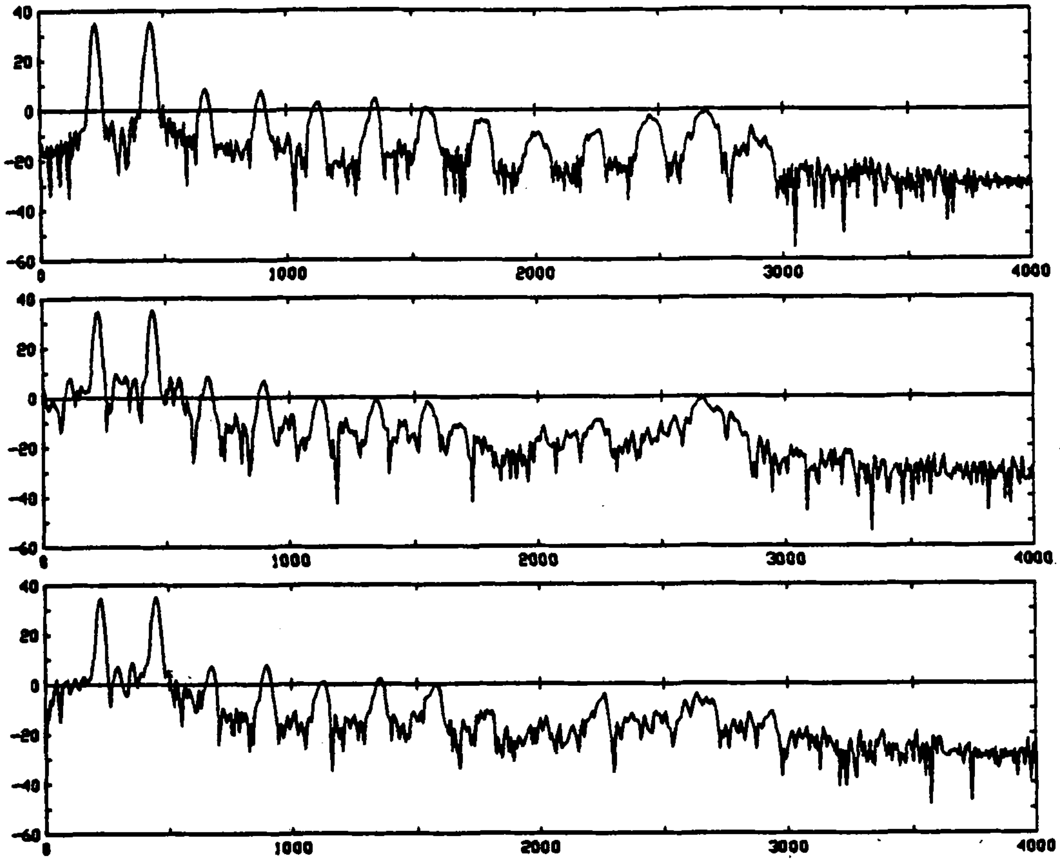

Fig. 9 Spectra for female speech.
(a) original speech,
(b) synthesized speech with the pitch lag constrained to be larger than 40 samples, and
(c) synthesized speech with the pitch lag allowed to fall below 40 samples.

8. Summary and Conclusions

The advantages of deriving the synthesis parameters based on generating the best frequency weighted error seem considerable. The sequential approach to choosing the pitch filter parameters is computationally attractive. In this approach, the pitch filter parameters are chosen with no input waveform. The pitch filter tries to generate an excitation waveform which is a scaled and delayed version of previous excitation waveform. The waveform selected from the dictionary then fills in the missing details. The approach taken allows the dictionary to have as few as 32 waveforms and still result in good quality output speech.

The synthesis parameters have been quantized at a rate corresponding to 4000 b/s. All indications are that a fully quantized 4800 b/s coder with relatively high quality speech is attainable with a computational complexity that is practical.

Appendix A. Error Minimization Model

Consider the analysis model shown in Fig. A.1. This system forms a filtered version of the linear combination of M input signals. The model will be used to find the set of coefficients $\{a_1, a_2, \ldots, a_M\}$ which minimize a squared error criterion. The model is general enough to subsume different types of analyses (covariance, windowed covariance and autocorrelation) through appropriate choices of the parameters. The model also can accommodate input signals which are time shifts of a basic signal.

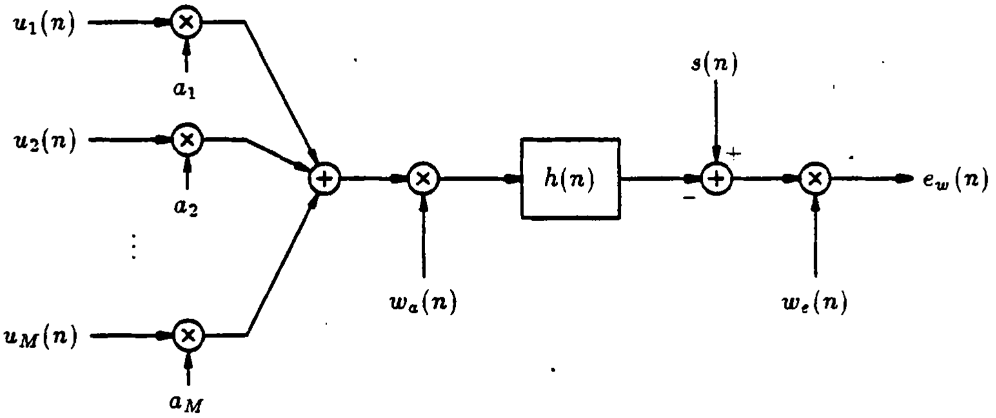

Fig. A.1 Error minimization model

In a block based analysis, a set of coefficients is determined for each block of N samples. The model includes time-windowing operations to localize the effect of the choice of the coefficients to the neighbourhood of a single block of samples. First the linear combination of the input signals is multiplied by a time window $w_a(n)$. The windowed sum passes through a weighting filter with impulse response $h(n)$. The error between a given reference signal $s(n)$ and the output of the weighting filter is formed. Finally, the error is windowed by an error window $w_e(n)$. Though not shown explicitly, the input signals themselves can also be windowed.

Conceptually the window $w_a(n)$ and the filter $h(n)$ in Fig. A.1 can be repeated in each of the input lines as shown in Fig. A.2. This leads to the equivalent problem of minimizing difference between $s(n)$ and the linear combination of the signals $v_i(n)$, where $v_i(n)$ is a filtered version of the input signal $u_i(n)$,

$$v_i(n) = \sum_{k=-\infty}^{\infty} w_a(k) u_i(k) h(n-k) . \tag{A.1}$$

The windowed error can be written as

$$e_w(n) = w_e(n)\Big[s(n) - \sum_{i=0}^{M} a_i v_i(n)\Big] . \tag{A.2}$$

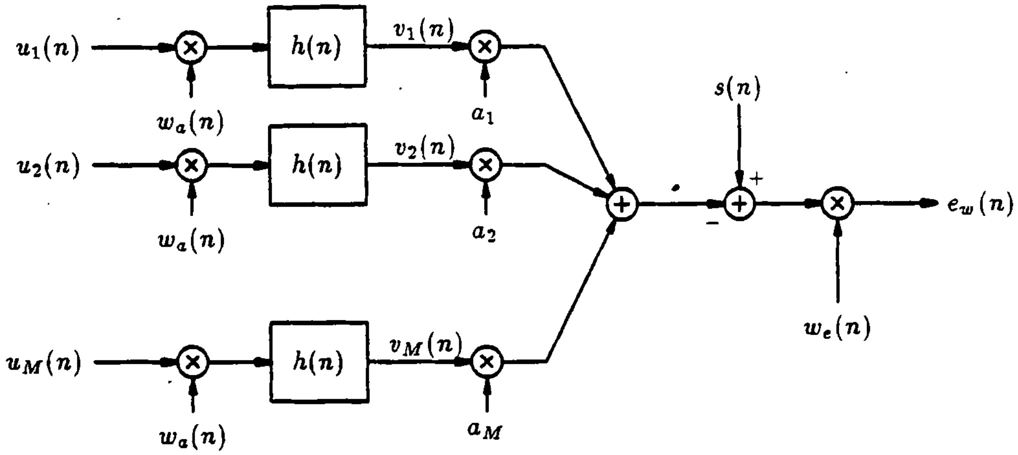

Fig. A.2 Modified error minimization model

The reference signal $s(n)$ is assumed to include components which compensate for the output due to other blocks of samples. This allows the analysis to proceed with zero initial conditions for the weighting filter.

The error criterion which is to be minimized is given as

$$\varepsilon = \sum_{n=-\infty}^{\infty} e_w^2(n) \,. \tag{A.3}$$

It is ε will be minimized with respect to the choice of the weights $\{a_1, a_2, \ldots, a_M\}$. The minimization can be carried out by differentiating ε with respect to each of the weights. This leads to a set of simultaneous equations which can be written in matrix form as $\mathbf{\Phi a} = \mathbf{b}$. The matrix of autocorrelation terms $\mathbf{\Phi}$ is given by

$$\mathbf{\Phi} = \sum_{n=-\infty}^{\infty} w_e^2(n)\mathbf{v}^{(n)}\mathbf{v}^{(n)^T} \,, \tag{A.4}$$

where the vector $\mathbf{v}^{(n)}$ is defined as

$$\mathbf{v}^{(n)} = \begin{bmatrix} v_1(n) \\ v_2(n) \\ \vdots \\ v_M(n) \end{bmatrix} . \tag{A.5}$$

The vector of coefficients $\mathbf{a}$ is defined as

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_M \end{bmatrix} . \tag{A.6}$$

The vector of cross-correlations $\mathbf{b}$ is defined as

$$\mathbf{b} = \sum_{n=-\infty}^{\infty} w_e^2(n)s(n)\mathbf{v}^{(n)} \,, \tag{A.7}$$

With the optimal choice of **a**, the resulting squared-error is

$$\varepsilon_{\mathrm{opt}} = \sum_{n=-\infty}^{\infty} w_e^2(n)s^2(n) - \mathbf{a}^T\mathbf{b} \ . \tag{A.8}$$

The last term is the decrease in squared-error due to the use of the optimal coefficients. There are several equivalent forms for this term,

$$\mathbf{a}^T\mathbf{b} = \mathbf{b}^T\Phi^{-1}\mathbf{b} = \mathbf{a}^T\Phi\mathbf{a} \ . \tag{A.9}$$

These expressions are valid only for the optimal choice of coefficients.

A.1 Covariance analysis

A covariance analysis is appropriate for a blockwise optimization. The coefficients will be chosen to minimize the sum of the squared error terms for a finite time interval. Consider the time-limited error window,

$$w_e(n) = \begin{cases} 1 & \text{for } 0 \le n < N \\ 0 & \text{elsewhere .} \end{cases} \tag{A.10}$$

For a block based analysis, it is also appropriate to have the coefficients $\{a_1, a_2, \ldots, a_M\}$ only apply to the portion of the input signal within the limits of the block. The effect of the input signals outside the block limits is absorbed into the "desired" signal $s(n)$. The window $w_a(n)$ is set to be the same as the error window. The last assumption is that the weighting filter is causal.

For the case just specified, the autocorrelation terms in the matrix become

$$\phi(i,j) = \sum_{n=0}^{N-1} v_i(n)v_j(n) \ . \tag{A.11}$$

The filtered signal $v_i(n)$ becomes

$$v_i(n) = \sum_{k=0}^{n} u_i(k)h(n-k) \ , \tag{A.12}$$

and the elements of the cross-correlation vector become

$$b(j) = \sum_{n=0}^{N-1} s(n)v_j(n) \ . \tag{A.13}$$

References

1. P. Kabal, "Code Excited Linear Prediction Coding of Speech at 4.8 kb/s", INRS-Telecommunications Technical Report, 1987.
2. C.C. Chu and P. Kabal, "Coding of LPC Parameters for Low Bit Rate Speech Coders", INRS-Telecommunications Technical Report 87-19, March 1987.
3. P. Kabal and R.P. Ramachandran, "Pitch prediction filters in speech coding", *submitted to IEEE Trans. Acoust., Speech, Signal Processing*, May 1987.