Regular-Pulse Excitation—A Novel Approach to Effective and Efficient Multipulse Coding of Speech

PETER KROON, STUDENT MEMBER, IEEE, ED F. DEPRETTERE, MEMBER, IEEE, AND ROB J. SLUYTER

Abstract—**This paper describes an effective and efficient time domain speech encoding technique that has an appealing low complexity, and produces toll quality speech at rates below 16 kbits/s. The proposed coder uses linear predictive techniques to remove the short-time correlation in the speech signal. The remaining (residual) information is then modeled by a low bit rate reduced excitation sequence that, when applied to the time-varying model filter, produces a signal that is "close" to the reference speech signal. The procedure for finding the optimal constrained excitation signal incorporates the solution of a few strongly coupled sets of linear equations and is of moderate complexity compared to competing coding systems such as adaptive transform coding and multipulse excitation coding. The paper describes the novel coding idea and the procedure for finding the excitation sequence. We then show that the coding procedure can be considered as an "optimized" baseband coder with spectral folding as high-frequency regeneration technique. The effect of various analysis parameters on the quality of the reconstructed speech is investigated using both objective and subjective tests. Further, modifications of the basic algorithm, and their impact on both the quality of the reconstructed speech signal and the complexity of the encoding algorithm, are discussed. Using the generalized baseband coder formulation, we demonstrate that under reasonable assumptions concerning the weighting filter, an attractive low-complexity/high-quality coder can be obtained.**

I. Introduction

An interesting application area for digital speech coding can be found in mobile telephony systems and computer networks. For these applications, toll quality speech at bit rates below 16 kbits/s is a prerequisite. Many of the conventional speech coding techniques [1] fail to obey this condition. However, a class of coders, the so-called delayed decision coders (DDC) [1, ch. 9], seems to be promising for these applications. Coders that belong to this class utilize an encoding delay to find the "best" quantized version of the input speech signal or a transformed version of it. Quite effective algorithms can be designed by combining predictive and DDC techniques to yield low bit rate waveform matching encoding schemes. A powerful and common approach is to use a slowly time-varying linear predictive (LP) filter to model the short-time spectral envelope of the quasi-stationary speech signal. The problem that remains is how to describe the resulting prediction residual that contains the necessary information to describe the fine structure of the underlying spectrum. In other words, what is the "best" low-capacity model for the speech prediction residual subjected to one or more judgment criteria. These may include objective and subjective quality measures (such as rate distortions and listening scores, respectively), but coder complexity can also be taken into account. Although certain models have been shown to behave very satisfactorily [2]–[4], the question of optimality remains difficult to answer.

In this paper we address the problem of finding an excitation signal for an LP speech coder that not only ensures a comparable quality with existing approaches, but is also structurally powerful. By the latter we mean that a fast realization algorithm and a corresponding high throughput (VLSI) implementation can be obtained. We propose a method in which the prediction residual is modeled by a signal that resembles an upsampled sequence and has, therefore, a regular (in time) structure. Because of this regularity, we refer to this coder as the regular-pulse excitation (RPE) coder [5]. The values of the nonzero samples in this signal are optimally determined by a least-squares analysis-by-synthesis fitting procedure that can be expressed in terms of matrix arithmetic.

In Section II we describe in more detail the regular-pulse excitation coding procedure and the algorithm for finding the excitation sequence. In Section III we show that the proposed encoding procedure can be interpreted in terms of optimized baseband coding. In Section IV, the influence of the various analysis parameters on the quality of the reconstructed speech is investigated. Further, to exploit the long-term correlation in the speech signal, the use of a pitch predictor is discussed. Modifications to the basic procedure, to attain a further reduction in complexity without noticeable quality loss, are described in Section V. Finally, in Section VI, we describe the effect of quantization on the quality of the reconstructed speech signal.

II. Basic Coder Structure

The basic coder structure can be viewed as a residual modeling process, as depicted in Fig. 1. In this figure, the residual $r(n)$ is obtained by filtering the speech signal $s(n)$

Manuscript received August 23, 1985; revised March 5, 1986. This work was supported in part by Philips Research Laboratories, Eindhoven, The Netherlands, and by the Dutch National Applied Science Foundation under Grant STW DEL 44.0643.

P. Kroon was with the Department of Electrical Engineering, Delft University of Technology, Delft, The Netherlands. He is now with the Acoustics Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974.

E. F. Deprettere is with the Department of Electrical Engineering, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands.

R. J. Sluyter is with the Philips Research Laboratories, 5600 MD Eindhoven, The Netherlands.

IEEE Log Number 8609633.

0096-3518/86/1000-1054$01.00 © 1986 IEEE

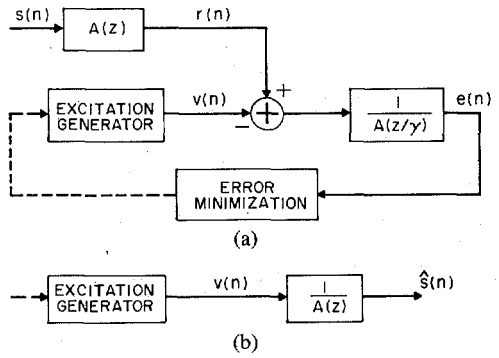

Fig. 1. Block diagram of the regular-pulse excitation coder: (a) encoder, (b) decoder.

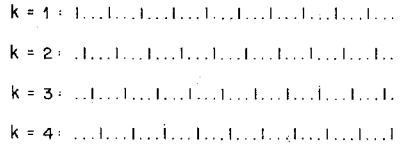

Fig. 2. Possible excitation patterns with $L = 40$ and $N = 4$.

through a pth-order time-varying filter $A(z)$,

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}, \tag{1}$$

which can be determined with the use of linear prediction (LP) techniques as described in, e.g., [6]. The difference between the LP-residual $r(n)$ and a certain model residual $v(n)$ (to be defined below) is fed through the shaping filter $1/A(z/\gamma)$,

$$\frac{1}{A(z/\gamma)} = \frac{1}{1 + \sum_{k=1}^{p} a_k \gamma^k z^{-k}}, \quad 0 \le \gamma \le 1. \tag{2}$$

This filter, which serves as an error weighting function, plays the same role as the feedback filter in adaptive predictive coding with noise shaping (APC-NS) [7] and the weighting filter in multipulse excitation (MPE) coders [2]. The resulting weighted difference $e(n)$ is squared and accumulated, and is used as a measure for determining the effectiveness of the presumed model $v(n)$ of the residual $r(n)$.

The excitation sequence $v(n)$ is determined for adjacent frames consisting of L samples each, and is constrained as follows. Within a frame, it is required to correspond to an upsampled version of a certain "optimal" vector $\boldsymbol{b} = (b(1), \cdots, b(Q))$ of length $Q (Q < L)$. Thus, each segment of the excitation signal contains Q equidistant samples of nonzero amplitude, while the remaining samples are equal to zero. The spacing between nonzero samples is $N = L/Q$. For a particular coder, the parameters L and N are optimally chosen but are otherwise fixed quantities. The duration of a frame of size L is typically 5 ms. Each excitation frame can support N sets of Q equidistant nonzero samples, resulting in N candidate excitation sequences. Fig. 2 shows the possible excitation patterns for a frame containing 40 samples and a spacing of $N = 4$. In this figure, the locations of the pulses are marked by a vertical dash and the zero samples by dots. If k ($k = 1, 2, \cdots, N$) denotes the *phase* of the upsampled version of the vector $\boldsymbol{b}^{(k)}$, i.e., the position of the first nonzero sample in a particular segment, then we have to compute for every value of k the amplitudes $b^{(k)}(\cdot)$ that minimize the accumulated squared error. The vector that yields the minimum error is selected and transmitted. The decoding procedure is then straightforward, as is shown in Fig. 1(b).

A. Encoding Algorithm

Denoting by $\boldsymbol{M}_k$ the Q by L position matrix with entries

$$\begin{aligned} &m_{ij} = 1 \quad \text{if } j = i*N + k - 1 \\ &m_{ij} = 0 \quad \text{otherwise} \end{aligned} \qquad \begin{aligned} &0 \le i \le Q - 1 \\ &0 \le j \le L - 1, \end{aligned} \tag{3}$$

the segmental excitation row vector $\boldsymbol{v}^{(k)}$, corresponding to the kth excitation pattern, can be written as

$$\boldsymbol{v}^{(k)} = \boldsymbol{b}^{(k)} \boldsymbol{M}_k. \tag{4}$$

Let $\boldsymbol{H}$ be an uppertriangular L by L matrix whose jth row ($j = 0, \cdots, L - 1$) contains the (truncated) response $h(n)$ of the error weighting filter $1/A(z/\gamma)$ caused by a unit impulse $\delta(n - j)$. That is,

$$\boldsymbol{H} = \begin{bmatrix} h(0) & h(1) & \cdots & h(L-1) \\ 0 & h(0) & & h(L-2) \\ 0 & 0 & & h(L-3) \\ \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & h(0) \end{bmatrix}. \tag{5}$$

If $\boldsymbol{e}_0$ denotes the output of the weighting filter due to the memory hangover (i.e., the output as a result of the initial filter state) of previous intervals, then the signal $e(n)$ produced by the input vector $\boldsymbol{b}^{(k)}$ can be described as

$$\boldsymbol{e}^{(k)} = \boldsymbol{e}^{(0)} - \boldsymbol{b}^{(k)} \boldsymbol{H}_k, \quad k = 1, \cdots, N, \tag{6}$$

where

$$\boldsymbol{e}^{(0)} = \boldsymbol{e}_0 + \boldsymbol{r}\boldsymbol{H}, \tag{7}$$

$$\boldsymbol{H}_k = \boldsymbol{M}_k \boldsymbol{H}, \tag{8}$$

and the vector $\boldsymbol{r}$ represents the residual $r(n)$ for the current frame. The objective is to minimize the squared error

$$E^{(k)} = \boldsymbol{e}^{(k)} \boldsymbol{e}^{(k)t}, \tag{9}$$

where t denotes transpose. For a given phase the optimal amplitudes $b^{(k)}(\cdot)$ can be computed from (6) and (9), by requiring $\boldsymbol{e}^{(k)} \boldsymbol{H}_k^t$ to be equal to zero. Hence,

$$\boldsymbol{b}^{(k)} = \boldsymbol{e}^{(0)} \boldsymbol{H}_k^t [\boldsymbol{H}_k \boldsymbol{H}_k^t]^{-1}. \tag{10}$$

By substituting (10) in (6) and thereafter the resulting expression in (9), we obtain the following expression for the error:

$$E^{(k)} = \boldsymbol{e}^{(0)} [\boldsymbol{I} - \boldsymbol{H}_k^t [\boldsymbol{H}_k \boldsymbol{H}_k^t]^{-1} \boldsymbol{H}_k] \, \boldsymbol{e}^{(0)t}. \tag{11}$$

The vector $\boldsymbol{b}^{(k)}$ that yields the minimum value of $E^{(k)}$ over all k is then selected. The resulting optimal excitation vector $\boldsymbol{v}^{(k)}$ is entirely characterized by its phase k and the corresponding amplitude vector $\boldsymbol{b}^{(k)}$. The whole procedure comprises the solution of N sets of linear equations as given by (10). A fast algorithm to compute the N vectors $\boldsymbol{b}^{(k)}$ simultaneously has been presented in [8] and [9]. We shall show in Section V that a further reduction in complexity can be obtained by exploiting the nature of the matrix product $\boldsymbol{H}_k\boldsymbol{H}_k^t$ in (10).

III. Generalized Baseband Coding

It may be observed that the regular-pulse excitation sequence bears some resemblance to the excitation signal of excited baseband coder (BBC) using spectral folding as high-frequency regeneration technique [4], [10]. In this section we show that the RPE coder can be interpreted as a generalized version of this baseband coder. For this purpose we use the block diagram of Fig. 3. The blocks drawn with solid lines represent the conceptual structure of a residual excited BBC coder with spectral folding. For this coder, the index k has no significance and is set to zero. In this scheme, the LP-residual signal $r(n)$, obtained by filtering the speech signal through the filter $A(z)$, is band-limited by an (almost) ideal low-pass filter $F_0(z)$, downsampled to $b^{(0)}(n)$ and transmitted. At the receiver, this signal is upsampled to $v^{(0)}(n)$ to recover the original bandwidth, and is fed through the synthesis filter to retrieve the speech signal $\hat{s}(n)$. When the dashed blocks are included in Fig. 3, one provides a possibility to optimize the filter $F_k(z)$, i.e., to replace the ideal low-pass filter $F_0(z)$ by another filter, which is more tailored to "optimal" waveform matching, where the optimality criterion is to minimize the (weighted) mean-squared error between the original and the reconstructed signal.

We shall now show that for this "optimized" BBC version, the output of the filter $F_k(z)$, after down- and upsampling, is exactly the excitation signal $v^{(k)}(n)$ as computed by the RPE algorithm. Thus, let there exist for each k, ($k = 1, \cdots, N$), an FIR filter $F_k(z)$ such that the weighted least-squares error $\Sigma_n e^2(n)$ over the interval L is minimal. Define $F_k(z)$ as

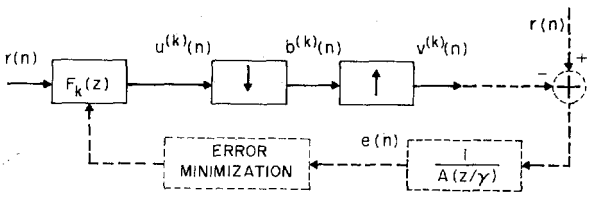

Fig. 3. Block diagram of a BBC coder (solid lines), and an RPE coder (solid and dashed lines).

$$F_k(z) = \sum_{i=0}^{L-1} f_i^{(k)} z^{-i}, \tag{12}$$

and

$$\boldsymbol{f}^{(k)} = [f^{(k)}(0)\, f^{(k)}(1) \cdots f^{(k)}(L-1)]. \tag{13}$$

Let $r_+(n)$ and $r_-(n)$ ($n = 0, \cdots, L-1$) denote the residual samples of the current frame and those of the previous frame, respectively. Then we can write for the output $u^{(k)}(n)$ of the filter $F_k(z)$

$$\boldsymbol{u}^{(k)} = \boldsymbol{f}^{(k)} \begin{bmatrix} r_+(0) & r_+(1) & \cdots & r_+(L-1) \\ r_-(L-1) & r_+(0) & & r_+(L-2) \\ r_-(L-2) & r_-(L-1) & & r_+(L-3) \\ \vdots & \vdots & & \vdots \\ r_-(1) & r_-(2) & & r_+(0) \end{bmatrix} = \boldsymbol{f}^{(k)}\boldsymbol{R}. \tag{14}$$

The vector $\boldsymbol{b}^{(k)}$, which is the downsampled version of $\boldsymbol{u}^{(k)}$ (with downsampling factor N), can be written as

$$\boldsymbol{b}^{(k)} = \boldsymbol{f}^{(k)}\boldsymbol{R}\boldsymbol{M}_k^t, \qquad = \boldsymbol{f}^{(k)}\boldsymbol{R}_k \tag{15}$$

with

$$\boldsymbol{R}_k = \begin{bmatrix} r_+(k-1) & r_+(N-1+k) & \cdots & r_+((Q-1)N-1+k) \\ r_-(L-2+k) & r_+(N-2+k) & & r_+((Q-1)N-2+k) \\ r_-(L-3+k) & & & \\ \vdots & \vdots & & \vdots \\ r_-(k) & & & r_-((Q-1)N+k) \end{bmatrix}, \tag{16}$$

where $\boldsymbol{M}_k$ is the position matrix as defined in (3), and where the definition $r_-(L+k) = r_+(k)$. The excitation vector $\boldsymbol{v}^{(k)}$ can be expressed as the product

$$\boldsymbol{v}^{(k)} = \boldsymbol{f}^{(k)} \begin{bmatrix} 0 \cdots 0 & r_+(k-1) & 0 \cdots 0 \cdots & r_+((Q-1)N-1+k) & 0 \cdots 0 \\ 0 \cdots 0 & r_-(L-2+k) & 0 \cdots 0 \cdots & r_+((Q-1)N-2+k) & 0 \cdots 0 \\ \cdots & \vdots & \cdots & \vdots & \cdots \\ \underbrace{0 \cdots 0}_{k-1} & r_-(k) & \underbrace{0 \cdots 0}_{N-k} & & \underbrace{0 \cdots 0}_{N-k} \end{bmatrix}. \tag{17}$$

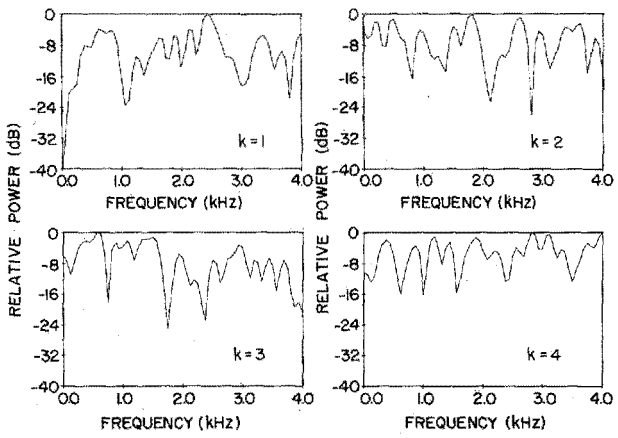

Fig. 4. Power spectra $|F_k(e^{j\theta})|^2$ for different values of k, obtained from a 5 ms speech segment.

Hence, with the matrix $\boldsymbol{H}$ and the initial error $\boldsymbol{e}^{(0)}$ as defined in the previous section,

$$\begin{aligned}\boldsymbol{e}^{(k)} &= \boldsymbol{e}^{(0)} - f^{(k)}\boldsymbol{R}_k\boldsymbol{M}_k\boldsymbol{H} \\ &= \boldsymbol{e}^{(0)} - f^{(k)}\boldsymbol{R}_k\boldsymbol{H}_k. \end{aligned} \tag{18}$$

Minimizing $\boldsymbol{e}^{(k)}\boldsymbol{e}^{(k)t}$, we obtain as solution

$$f^{(k)} = \boldsymbol{e}^{(0)}(\boldsymbol{R}_k\boldsymbol{H}_k)^t[\boldsymbol{R}_k\boldsymbol{H}_k(\boldsymbol{R}_k\boldsymbol{H}_k)^t]^{-1}. \tag{19}$$

Substituting this result in (15), we obtain the vector $\boldsymbol{b}^{(k)}$, which is equal to the pulse amplitude vector $\boldsymbol{b}^{(k)}$ obtained via the procedure described in Section II (see the proof in the Appendix).

Fig. 4 gives an example of the spectra $|F_k(e^{j\theta})|^2$ obtained from real speech data. From this figure we see that the filters $F_k(z)$ are rather different from the one ($F_0(z)$) used in the classical baseband coder, and have a more all-pass character.

Although the RPE algorithm and the optimal BBC algorithm are conceptually equivalent, the optimized BBC variant will in general not offer any computational advantage over the RPE approach. However, in Section V, it is demonstrated that under certain reasonable assumptions concerning the weighting filter, the BBC approach can provide an attractive alternative in practice.

IV. Evaluation of the RPE Algorithm

Fig. 5 shows a typical example of the waveforms as produced by the RPE coder, using the analysis parameters listed in Table I. The corresponding short-time power spectra of the speech signal $s(n)$ (solid line) and the reconstructed signal $\hat{s}(n)$ (dashed line) are shown in Fig. 6. To give an impression of the signal-to-noise ratio over a complete utterance, we show in Fig. 7 the segmental SNR (SNRSEG) computed every 10 ms for the utterance "a lathe is a big tool" spoken by both a female and a male speaker.

A. RPE Analysis Parameters

The RPE analysis parameters that could affect the final speech quality are listed below:

1) predictor parameters,

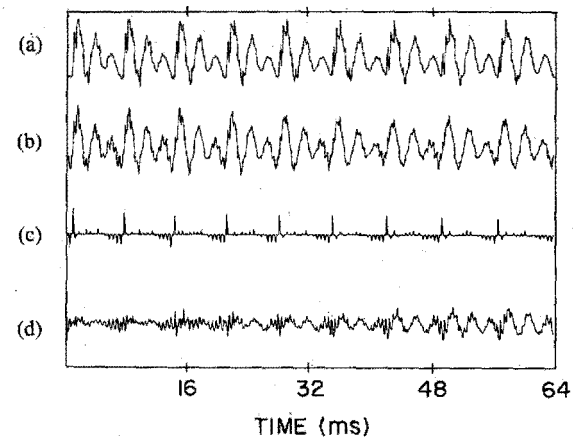

Fig. 5. (a) Speech signal $s(n)$, (b) reconstructed speech signal $\hat{s}(n)$, (c) excitation signal $v(n)$, and (d) difference signal $s(n) - \hat{s}(n)$ in the RPE coding procedure.

TABLE I
Default Parameters RPE Analysis

Parameter	Value
sampling frequency	8 kHz
LP analysis procedure	autocorrelation
order (p)	12
update rate coefficients	10 ms
analysis frame size	25 ms Hamming window
pulse spacing N	4
frane size L	5 ms
weight factor γ	0.80

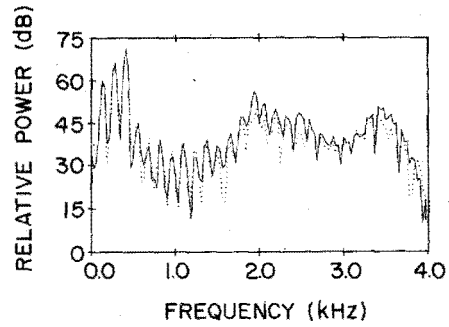

Fig. 6. Power spectra of the original speech segment (solid line) and the reconstructed speech segment (dashed line). The spectra were obtained with a Hamming window using the last 32 ms segment of the data displayed in Fig. 5.

2) pulse spacing N,
3) frame size L, and
4) error weighting filter.

To evaluate the coder behavior, we used a set of default parameter values (see Table I), while the parameter under investigation was varied.

The effects of the predictor parameters in APC-like schemes have been extensively studied in the literature (e.g., [1]), and will not be discussed in detail in this paper. We found that good results were obtained with the autocorrelation method using a Hamming window on 25 ms frames. The predictor coefficients were updated every 20 ms and the predictor order p was chosen to be equal to 12.

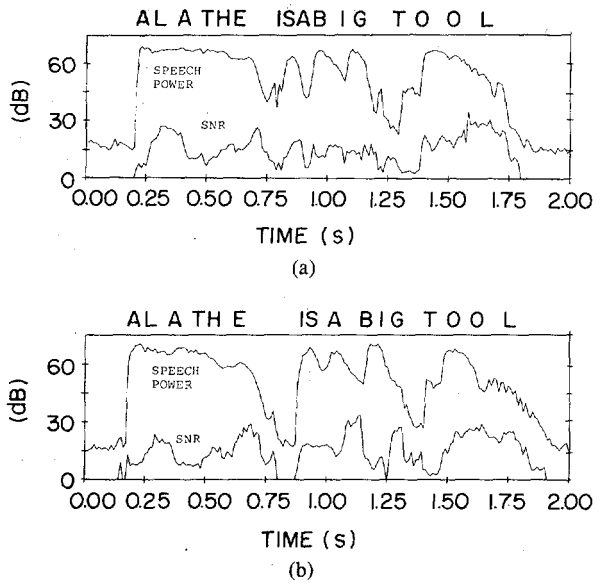

Fig. 7. Segmental SNR for successive time frames for a female speaker (a) and a male speaker (b). The upper curve represents the speech power +15 dB.

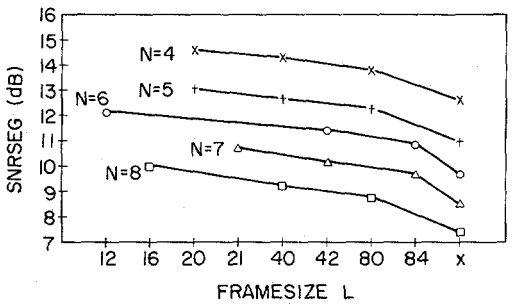

Fig. 8. Segmental SNR values for different frame sizes L and pulse spacings N. The results for $L = X$ were obtained with a fixed value for k and $L = 40$.

We mentioned earlier that for the case in which there is no phase adaptation (on a frame basis), that is, k is fixed and equal to 1, the structure of the excitation signal resembles the upsampled residual signals used in BBC coders with spectral folding. This observation can give us a rough estimate of the maximum spacing (N) between the pulses, to ensure a good synthetic speech quality. Assuming a maximum fundamental frequency of 500 Hz, we have to use a sampling rate of minimally 1000 Hz. Hence, for an 8 kHz sampling rate, the pulse spacing should be less than or equal to 8.

To investigate the effect of different frame sizes L and pulse spacings N, we computed the segmental SNR values of the reconstructed speech signals for various values of these parameters. Fig. 8 shows the averaged segmental SNR values for two female and two male speakers[1] for different values of N and L. As far as possible, we have chosen the same frame size for different values of N. From this figure, we see that the SNR increases with the number of pulses and decreases with increasing frame size. However, there is no real tradeoff between the values of L and N. Informal listening tests confirmed the ranking as introduced by the SNR measurements. For values of N greater than 5, some of the utterances (especially those by female speakers) sounded distorted. From our experiments, we found that $N = 4$ and $L = 5$ ms will give the best results considering the bit rate constraints.

[1]The utterances are: "A lathe is a big tool" and "An icy wind raked the beach."

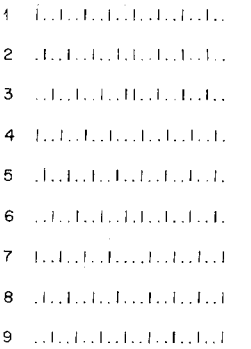

Fig. 9. RPE excitation patterns with $D = 12$, $L = 24$, and $N = 3$.

TABLE II
SNR VALUES FOR DIFFERENT VALUES OF L AND D

$L:D$	SNRSEG	SNR
20:20	14.58 dB	11.80 dB
40:20	14.75 dB	11.90 dB
40:40	14.28 dB	11.17 dB
80:40	14.58 dB	11.29 dB
80:80	13.80 dB	10.44 dB

The pulse amplitudes $b^{(k)}(\cdot)$ and phase k are computed every L samples, which means that the *phase adaptation rate* is equal to $1/L$. To investigate the effect of this "disturbance," without changing the size of L, we considered phase adaptation every D samples, where the value of D is less than or equal to L, and L/D must be an integer ratio. Within a frame of size L, the possible number of excitation sequences is then given by

$$B = N^d, \quad d = L/D. \tag{20}$$

Hence, a value of D smaller than L results in a more complex procedure for the computation of the optimum excitation. Fig. 9 shows the possible excitation patterns for $L = 24$, $D = 12$, and $N = 3$. Table II lists the resulting averaged SNR values for different frame sizes L and ratios $L/D = 1$ and 2. From this table we see a small improvement in SNR for values of D less than L, at the expense of a much higher complexity.

B. Application of a Pitch Predictor

An examination of the regular-pulse excitation (see, for example, Fig. 5) reveals the periodic structure of the excitation for voiced sounds. Obviously, the RPE algorithm aligns the excitation "grid" to the major pitch pulses, thereby introducing the possibility that the remaining pulses within the grid are not optimally located. If we model the major pitch pulses with a pitch predictor/synthesizer, the remaining excitation sequence can be mod-

eled by the regular-pulse excitation sequence. A simple but effective pitch predictor is the so-called one-tap predictor,

$$1 - P(z) = \beta z^{-M}, \tag{21}$$

where M represents the distance between adjacent pitch pulses and β is a gain factor. The pitch predictor parameters can be determined either in an open-loop configuration [11], or in a closed-loop configuration [12]. In the latter case, the parameters can be optimally computed by including a pitch generator $1/P(z)$ in the closed-loop diagram of Fig. 1. The parameters β and M are determined such that the output of the pitch generator due to its initial state is optimally close (in the weighted sense) to the initial error signal $e^{(0)}(n)$. Once β and M have been determined, the remaining regular-pulse excitation signal is computed as described in Section II, except that this signal is now to be fed through both the pitch generator and the weighting filter. The advantage of determining the pitch parameters within the analysis loop is that the pitch generator is then optimally contributing to the minimization of the weighted error. To be more specific, let $y_M(n)$ be the response of the pitch generator to an input $v(n)$, which is zero for $n \geq 0$,

$$y_M(n) = v(n) + \beta y_M(n - M). \tag{22}$$

Let $z_M(n)$ represent the response of the weighting filter to the input signal $y_M(n)$, defined in (22), and let $e^{(0)}(n)$ represent the initial error as defined in (7). The error to be minimized will then be

$$E(M, \beta) = \sum_n (e^{(0)}(n) - \beta z_M(n))^2. \tag{23}$$

The approach is to compute β for all possible values of M within a specified range, and then select the pair (M, β) for which $E(M, \beta)$ is minimal.

The range of M should be chosen to accommodate to the variation in pitch frequency in the speech signal. However, in simulations with a one-tap predictor using different ranges of M, we found that a range of M between 16 and 80 (i.e., a fundamental frequency between 100 and 470 Hz) is satisfactory. The effect of pitch prediction is demonstrated in Fig. 10, by using the same speech segment as used in Fig. 5. The short-time power spectra of the speech signal $s(n)$ (solid line) and of the error signal $s(n) - \hat{s}(n)$ (dashed line) for $\gamma = 0.80$, without and with pitch filter, are shown in Figs. 11 and 12, respectively. The effect of pitch prediction on the averaged segmental SNR values is shown in Fig. 13. These figures show that the effect of pitch prediction is to decrease the absolute level of noise power and to flatten its spectrum, and thereby improving the performance in terms of SNR. This effect was most noticeable for high-pitched (average pitch $\geq$250 Hz) speakers.

C. Error Weighting Filter

Although the effect of noise shaping can be heard, the real mechanism behind this effect is not clear. We will not pursue the question whether the proposed noise-shap-

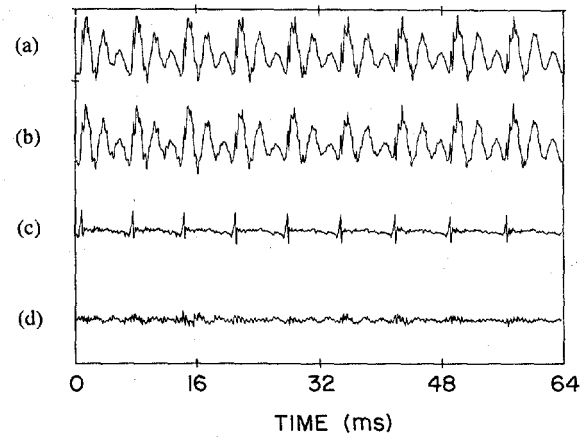

Fig. 10. (a) Speech signal $s(n)$, (b) reconstructed speech signal $\hat{s}(n)$, (c) excitation signal (i.e., output of the pitch generator), (d) difference signal $s(n) - \hat{s}(n)$ in the RPE coding procedure with pitch prediction.

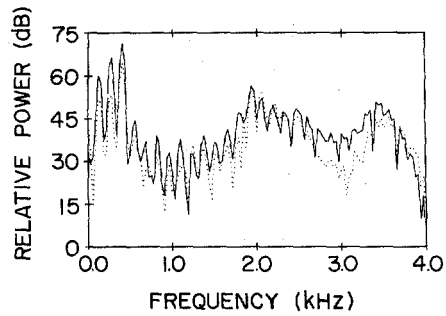

Fig. 11. Power spectra of the speech signal (solid line) and the difference signal $s(n) - \hat{s}(n)$ (dashed line) for $\gamma = 0.80$. The spectra were obtained from the last 32 ms segment of Fig. 5.

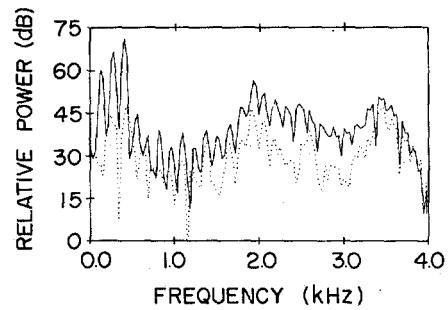

Fig. 12. Power spectra of the speech signal (solid line) and the difference signal $s(n) - \hat{s}(n)$ (dashed line) for $\gamma = 0.80$, and pitch prediction. The spectra were obtained from the last 32 ms segment of Fig. 10.

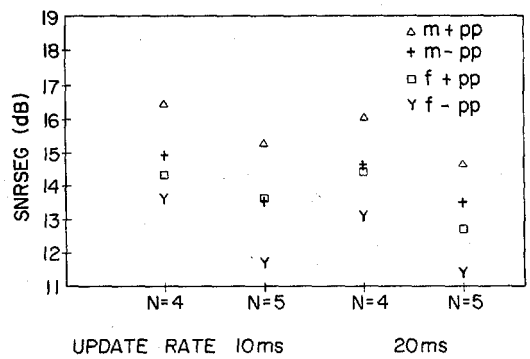

Fig. 13. Segmental SNR values obtained from RPE encoded speech with (+*pp*) and without (−*pp*) pitch prediction for different update rates of the predictors and different pulse spacings N (f = female, m = male).

ing filter of (2) is an effective choice or not, and concentrate instead on the effect of the suggested filter and its control parameter γ. This parameter determines the amount of noise power in the formant regions of the speech spectrum. Noise shaping reduces the SNR, but improves the perceived speech quality. An optimal value for γ was found to be between 0.80 and 0.90 at an 8 kHz sampling rate, and resulted in an average 2 dB decrease in SNR.

Aside from the value of γ, the order of the noise-shaping filter could also be of importance. By default, the coefficients $\{a_k\}$ and the order p of $1/A(z/\gamma)$ are equal to those of the predictor $A(z)$, but instead, we can compute a qth-order predictor ($q < p$) and use the resulting q coefficients to define the weighting filter. While reducing the order, we nevertheless must take care that the noise remains properly weighted. We examined the effect of decreasing the order of the weighting filter $1/A(z/\gamma)$, and observed that for low orders (2–4), the results were close to those obtained with a 16th-order filter. However, the computational savings obtained by reducing the order of the weighting filter are marginal.

The time-varying nature of the weighting filter provides a significant contribution to the complexity of the analysis procedure, since the system of linear equations to be solved is entirely built on the impulse response of this filter. It is obvious that the computational complexity would be considerably lower in case a weighting filter could be chosen such that the matrix to be inverted no longer depends on short-time data. It turns out that this is possible by choosing the weighting filter equal to $1/C(z/\gamma)$,

$$\frac{1}{C(z/\gamma)} = \frac{1}{1 + \sum_{k=1}^{9} c_k \gamma^k z^{-k}}, \tag{24}$$

where $\{c_k\}$ are the coefficients of the *fixed* low-order predictors as used in DPCM systems, which are based on the averaged spectral characteristics of speech. We carried out comparative listening tests on the results obtained with fixed weighting filters of different orders ($q = 1$ to 3). The value of γ was set to 0.80 and we used for $\{c_k\}$ the coefficients tabulated in [13]. It was surprising to find that the effects of the weighting filters $1/A(z/\gamma)$ and $1/C(z/\gamma)$ were judged to be almost equivalent. This remarkable result can be exploited to dramatically reduce the complexity of the proposed coder, as we will show in the next section.

V. Complexity Reduction of the RPE Coder

The analysis procedure of the RPE coder necessitates the solution of N sets of linear equations, where N represents the spacing between successive pulses within a frame in the excitation model. However, the matrices $\boldsymbol{H}_k\boldsymbol{H}_k^t$, which have to be inverted, can be solved very efficiently as was described in [8] and [9]. We shall not pursue the details of this procedure here, but we shall instead look for modifications of the algorithm to reduce the complexity without affecting the coder performance.

A. Modification of $\boldsymbol{H}_k\boldsymbol{H}_k^t$ to a Toeplitz Matrix

To begin with, we can reconfigure the algorithm to force the matrix product $\boldsymbol{H}_k\boldsymbol{H}_k^t$ in (10) to become a single Toeplitz matrix which is independent of the phase k. Thus, let

$$h(n) = \gamma^n g(n), \quad n = 0, 1, 2, \cdots, \tag{25}$$

be the impulse response of the weighting filter $1/A(z/\gamma)$, where $g(n)$ is the impulse response of the all-pole filter $1/A(z)$. For values of $|\gamma|$ less than one, $h(n)$ converges faster to zero than $g(n)$ and, as a result, the L by $2L$ matrix built on $h(n)$ can be very well approximated by the uppertriangular Toeplitz matrix $\boldsymbol{H}$ in (26).

$$\boldsymbol{H} = \begin{bmatrix} h(0) & h(1) & \cdots & h(L-1) & 0 & \cdots & 0 \\ 0 & h(0) & \cdots & h(L-2) & h(L-1) & & 0 \\ 0 & 0 & \cdots & h(L-3) & h(L-2) & \vdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & h(0) & \cdots & h(L-1) & 0 \end{bmatrix}. \tag{26}$$

Notice that the matrix $\boldsymbol{H}\boldsymbol{H}^t$ is also a Toeplitz matrix. Moreover, when substituting $\boldsymbol{H}$ from (26) into (8), we shall have that the matrices $\boldsymbol{H}_k\boldsymbol{H}_k^t$ are independent of the phase index k and are equal to a single Toeplitz matrix. It should also be remarked that the matrix of (26) is an L by $2L$ matrix instead of an L by L. Thus, when substituting $\boldsymbol{H}$ of (26) in (7) and (8), the vectors $\boldsymbol{e}_0$, $\boldsymbol{e}^{(0)}$, and $\boldsymbol{e}^{(k)}$ in (6) and (7) will now be of length $2L$, while the vectors $\boldsymbol{v}^{(k)}$ and $\boldsymbol{r}$ in (4) and (7) remain of dimension L. The RPE encoding procedure that is based on the mapping $\boldsymbol{H}$ in (26), and for which $g(n)$ in (25) is the impulse response of the transfer function $1/A(z)$, will be referred to as RPM1. Fig. 14(a) shows the segmental SNR values per 10 ms for this method (dashed line) and the original method (solid line) for the utterance "a lathe is a big tool" spoken by both a male and a female speaker.

B. Modification of $\boldsymbol{H}_k\boldsymbol{H}_k^t$ to a Band Matrix

In the previous subsection, a computationally attractive scheme was obtained by forcing the matrix operator $\boldsymbol{H}$ to be of the form of (26). Recall, however, that this structure is almost naturally emerging when the mapping originally defined via (5) is taken to be of dimension L by $2L$ instead of L by L. This is the more so when $h(n)$ in (26) is the impulse response of the fixed filter $1/C(z/\gamma)$ of (24). But an even more interesting observation is that the resulting single Toeplitz matrix, whether data dependent or not, is strongly diagonal dominant. Hence, when minimizing $E^{(k)}$ in (11), where now $\boldsymbol{H}$ is built on $1/C(z/\gamma)$, or equivalently, when maximizing

$$T^{(k)} = \boldsymbol{e}^{(0)}\boldsymbol{H}_k^t[\boldsymbol{H}_k\boldsymbol{H}_k^t]^{-1}\boldsymbol{H}_k\,\boldsymbol{e}^{(0)t}, \tag{27}$$

we can conveniently replace the (Toeplitz) matrix $\boldsymbol{H}_k\boldsymbol{H}_k^t$ with a diagonal matrix $r_0\boldsymbol{I}$, where $r_0 = \sum_{i=0}^{L-1} h^2(i)$, yield-

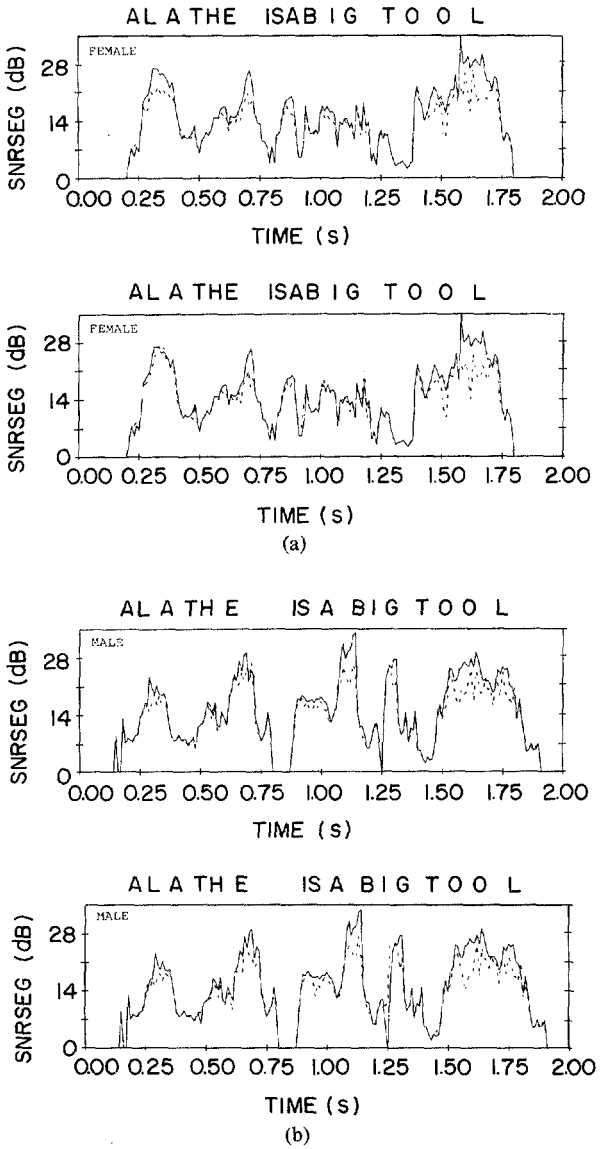

Fig. 14. Segmental SNR ratios for RPE (solid line) and modified methods, (a) RPM1 and (b) RPM2 (dashed line) for a female and male speaker.

TABLE III
SNR VALUES OBTAINED WITH THE ORIGINAL AND THE MODIFIED RPE ALGORITHMS RPM1 AND RPM2 DESCRIBED IN SECTION V-A AND V-B. THE PROCEDURES RPF1 AND RPF2 ARE DESCRIBED IN SECTION V-C.

Method	SNRSEG	SNR
RPE	14.28 dB	11.17 dB
RPM1	12.98 dB	10.93 dB
RPM2	13.00 dB	11.03 dB
RPF1	10.04 dB	9.38 dB
RPF2	10.40 dB	9.21 dB

ing

$$T^{(k)} = \frac{1}{r_0} e^{(0)} H_k^t H_k e^{(0)t}, \tag{28}$$

which means that no matrix inversion is needed to find the optimum phase k. Table III lists the SNR and SNRSEG values for the different methods, obtained by averaging the results of the same four utterances used in previous examples. Method RPM2 refers to the procedure described in this subsection, where the optimal phase index k is determined from (28), after which the excitation string $\boldsymbol{b}^{(k)}$ is computed according to (10). From this table, we see that the modifications introduced resulted in a slight decrease in SNR. But from informal listening tests, the modified methods were judged to be almost equivalent to the original RPE method. Fig. 14(b) shows the segmental SNR values per 10 ms for RPM2 (dashed lines) and the original method (solid line) for the utterance "a lathe is a big tool" spoken by both a male and a female speaker.

C. Avoiding Matrix Inversion

The discussions in the previous two subsections have led to the conclusion that the complexity of the RPE coder, although moderate by itself, can be substantially reduced without any significant degradation of the speech quality. We shall show in this subsection that it is even possible to obtain an extremely simple encoding algorithm that turns out to yield an applicable practical version of the (conceptual) optimal baseband coder which was described in Section III and was shown there to be equivalent to the RPE coder. Thus, let $h(n)$ in (26) be the impulse response of the time-invariant filter $1/C(z/\gamma)$ as defined in (24). Next, use in (8) the matrix $\boldsymbol{H}$ as defined in (26) and discard the zeroth-order approximation $\boldsymbol{e}_0$ in (7). Then (6) and (10) become

$$\boldsymbol{e}^{(k)} = \boldsymbol{r}\boldsymbol{H} - \boldsymbol{b}^{(k)}\boldsymbol{H}_k, \tag{29}$$

and

$$\boldsymbol{b}^{(k)}[\boldsymbol{H}_k\boldsymbol{H}_k^t] = \boldsymbol{r}\boldsymbol{H}\boldsymbol{H}^t\boldsymbol{M}_k^t, \tag{30}$$

respectively. Now denoting

$$\boldsymbol{S} = \boldsymbol{H}\boldsymbol{H}^t, \tag{31}$$

and recalling that

$$\boldsymbol{H}_k\boldsymbol{H}_k^t \approx r_0\boldsymbol{I}, \tag{32}$$

with

$$r_0 = \sum_{i=0}^{L-1} h^2(i),$$

as a coder constant, it is easy to show that

$$\boldsymbol{b}^{(k)} = \frac{1}{r_0}\boldsymbol{r}\boldsymbol{S}\boldsymbol{M}_k^t. \tag{33}$$

Interpreting $\boldsymbol{M}_k^t$ as a downsampling operator, (33) says that $\boldsymbol{b}^{(k)}$ resembles[2] a downsampled output of a smoother $\boldsymbol{S}$ whose input is a scaled version of the residual $\boldsymbol{r}$ [see Fig. 15(a)]. The excitation selection in the diagram of Fig. 15(a) is based on the minimization of the approximation error given by (11). Under the above-mentioned constraints, this equation becomes

$$E^{(k)} = \boldsymbol{r}\boldsymbol{H}\boldsymbol{H}^t\boldsymbol{r}^t - r_0\boldsymbol{b}^{(k)}\boldsymbol{b}^{(k)t}. \tag{34}$$

[2]This statement must be carefully interpreted. In fact, (33) is a block smoother, and hence, the boundary conditions of the smoother's internal state must be properly taken into account.

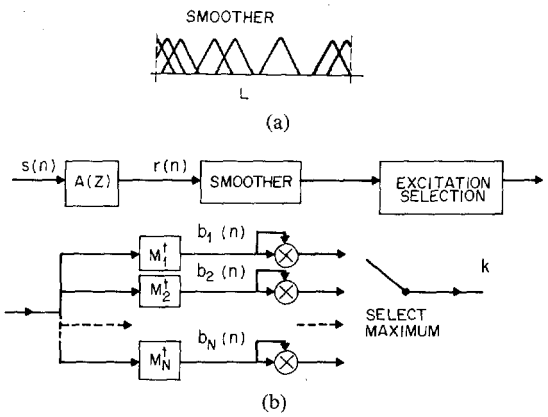

Fig. 15. Simplified RPE procedure (a) and excitation selection (b). The smoother is represented by a triangle shape.

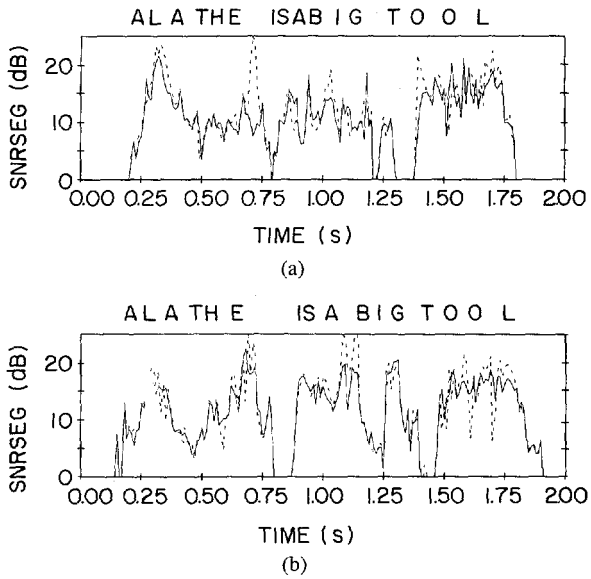

Fig. 16. Segmental SNR for RPF1 procedure (solid line) and RPF2 procedure (dashed line) for a female (a) and a male (b) speaker.

Hence,

$$\min \{E^{(k)}\} = \max \{\boldsymbol{b}^{(k)}\boldsymbol{b}^{(k)t}\}. \tag{35}$$

The whole procedure is now extremely simple. The residual signal $\boldsymbol{r}$ is "smoothed" with the smoother $\boldsymbol{S} = \boldsymbol{H}\boldsymbol{H}^t$. The resulting output vector is downsampled by applying $\boldsymbol{M}_k^t$, and the $\boldsymbol{b}^{(k)}$ for which $\boldsymbol{b}^{(k)}\boldsymbol{b}^{(k)t}$ is maximum is selected [see Fig. 15(b)]. Notice that since $\boldsymbol{H}$ is built on $1/C(z/\gamma)$, the smoother $\boldsymbol{S}$ will be of low order (typically 3rd order), since $h(n)$ is a rapidly decaying sequence. For comparison, the averaged SNRSEG values obtained with this procedure have been included in Table III. In this table, the RPE coder using a fixed weighting filter is referred to as RPF1, while the procedure outlined above is referred to as RPF2. In Fig. 16, the same comparison is made of the segmental SNR as a function of time for the utterance "a lathe is a big tool" spoken by both a male and a female speaker. From this figure, it is clear that for a fixed weighting filter procedure RPF2 provides a quality comparable to that of procedure RPF1. The advantage of the former is its ease of implementation.

VI. Quantization

To quantize the pulses (i.e., entries of $\boldsymbol{b}^{(k)}$), we used an 8-level adaptive quantizer whose input range was adjusted to the largest pulse amplitude within the current frame of size L. The quantization bins can be determined by a Lloyd–Max procedure (nonuniform), but we found that a uniform quantizer also performs quite well. The quantizer normalization factor is logarithmically encoded with 6 bits and is transmitted every L samples (typically 5 ms). The normalized pulses are encoded using 3 bits per pulse. To minimize quantization errors, the quantizer has to be incorporated in the minimization procedure. This can be done in two ways. In the first case (RPQ1), only the optimal excitation vector is quantized; and in the second case (RPQ2), every candidate $\boldsymbol{b}^{(k)}$ is quantized and the quantized vector that produces a minimum error is selected. From segmental SNR measurements, we found that RPQ2 yields a higher SNR, and in listening tests the quality of the reconstructed speech of RPQ2 was judged to be somewhat better than that of RPQ1.

The 12 reflection coefficients were transformed to inverse sine coefficients and encoded with 44 bits/set. The bit-allocation and quantizer characteristics were determined by the minimum deviation method [14]. Using 3 bits/pulse and a pulse spacing of $N = 4$, the excitation signal can be encoded with 7 kbits/s. The predictor coefficients can be encoded with 2.2 kbits/s resulting in a total bit rate of 9.2 kbits/s. The quality of the reconstructed speech was judged to be good but definitely not transparent. In informal listening tests, it was determined that the RPE approach has fewer artifacts than the baseband coder as proposed in [4], and that the performance is comparable to that of the MPE schemes. A pitch predictor will enhance the coder performance but goes at the cost of an additional 1000 bits/s (4 bits for β and 6 bits for M).

VII. Conclusion

In this paper, a novel coding concept has been proposed that uses linear prediction to remove the short-time correlation in the speech signal. The remaining residual signal is then modeled by a regular (in time) excitation sequence, that resembles an upsampled sequence. This model excitation signal is determined in such a way that the perceptual error between the original and the reconstructed signal is minimized. The computational effort is only moderate and can be further reduced by using a fixed error weighting filter and an appropriate vector size (minimization segment length). The coder can produce high-quality speech at bit rates around 9600 bits/s by using a pulse spacing equal to 4 and quantizing each pulse with 3 bits. The use of pitch prediction improves the speech quality but, in general, the RPE coder performs adequately without a pitch predictor. Other applications for the proposed coder can be found in the area of wide-band speech coding (7 kHz bandwidth) as encountered in tele- and video-conferencing applications [15].

APPENDIX

The excitation vector $\boldsymbol{b}^{(k)}$, obtained with the optimized BBC (Section III), coincides with the vector $\boldsymbol{b}^{(k)}$ produced by the RPE algorithm (Section II).

Proof: Equation (19) can be written as

$$\boldsymbol{f}^{(k)}\boldsymbol{R}_k[\boldsymbol{H}_k\boldsymbol{H}_k^t]\boldsymbol{R}_k^t = \boldsymbol{e}^{(0)}\boldsymbol{H}_k^t\boldsymbol{R}_k^t.$$

Multiplying both sides to the right by $\boldsymbol{R}_k$ gives

$$\boldsymbol{f}^{(k)}\boldsymbol{R}_k[\boldsymbol{H}_k\boldsymbol{H}_k^t]\boldsymbol{R}_k^t\boldsymbol{R}_k = \boldsymbol{e}^{(0)}\boldsymbol{H}_k^t\boldsymbol{R}_k^t\boldsymbol{R}_k.$$

Now assuming that $\boldsymbol{R}_k^t\boldsymbol{R}_k$ is nonsingular (which will almost always be the case for speech signals), we can as well write

$$\boldsymbol{f}^{(k)}\boldsymbol{R}_k[\boldsymbol{H}_k\boldsymbol{H}_k^t] = \boldsymbol{e}^{(0)}\boldsymbol{H}_k^t.$$

Substituting $\boldsymbol{b}^{(k)}$ for $\boldsymbol{f}^{(k)}\boldsymbol{R}_k$, see (15), in this equation, we obtain (10). □

REFERENCES

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

[2] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1982, pp. 614–617.

[3] 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*, 1985, pp. 937–940.

[4] R. J. Sluyter, G. J. Bosscha, and H. M. P. T. Schmitz, "A 9.6 kbit/s speech coder for mobile radio applications," in *Proc. IEEE Int. Conf. Commun.*, May 1984, pp. 1159–1162.

[5] E. F. Deprettere and P. Kroon, "Regular excitation reduction for effective and efficient LP-coding of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1985, pp. 25.8.1–25.8.4.

[6] J. D. Markel and A. H. Gray, *Linear Prediction of Speech.* Berlin, Germany: Springer-Verlag, 1976.

[7] B. S. Atal, "Predictive coding of speech at low bit rates," *IEEE Trans. Commun.*, vol. COM-30, pp. 600–614, Apr. 1982.

[8] E. F. Deprettere and K. Jainandunsing, "Design and VLSI implementation of a concurrent solver for *N*-coupled least-squares fitting problems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1985, pp. 6.3.1–6.3.4.

[9] K. Jainandunsing and E. F. Deprettere, "Design and VLSI implementation of a concurrent solver for *N*-coupled least-squares fitting problems," *IEEE J. Select. Areas Commun.*, pp. 39–48, Jan. 1986.

[10] V. R. Viswanathan, A. L. Higgins, and W. H. Russel, "Design of a robust baseband LPC coder for speech transmission over noisy channels," *IEEE Trans. Commun.*, vol. COM-30, pp. 663–673, Apr. 1982.

[11] P. Kroon and E. F. Deprettere, "Experimental evaluation of different approaches to the multi-pulse coder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1984, pp. 10.4.1–10.4.4.

[12] S. Singhal and B. S. Atal, "Improving performance of multipulse LPC coders at low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1984, pp. 1.3.1–1.3.4.

[13] J. L. Flanagan, M. R. Schroeder, B. S. Atal, R. E. Crochiere, N. S. Jayant, and J. M. Tribolet, "Speech coding," *IEEE Trans. Commun.*, vol. COM-27, pp. 710–736, Apr. 1979.

[14] A. H. Gray and J. D. Markel, "Implementation and comparison of two transformed reflection coefficient scalar quantization methods," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 575–583, Oct. 1980.

[15] C. Horne, P. Kroon, and E. F. Deprettere, "VLSI implementable algorithm for transparent coding of wide band speech below 32 kb/s," in *Proc. IASTED Symp. Appl. Signal Processing Dig. Filter.*, June 1985, pp. A3.1–A3.4.

Peter Kroon (S'82–M'86) was born in Vlaardingen, The Netherlands, on September 7, 1957. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from Delft University of Technology, Delft, The Netherlands, in 1978, 1981, and 1985, respectively.

His Ph.D. work focused on time-domain techniques for toll quality speech coding at rates below 16 kbits/s. From 1982 to 1983 he was a Research Assistant at the Network Theory Group, Delft University of Technology. During the years 1984 and 1985 he was sponsored by Philips Research Labs to work on coders suitable for mobile radio applications. He is currently with the Acoustics Research Department, AT&T Bell Laboratories, Murray Hill, NJ. His research interests include speech coding, signal processing, and the development of software for signal processing.

Ed F. Deprettere (M'83) was born in Roeselare, Belgium, on August 10, 1944. He received the M.S. degree from the Gent State University, Gent, Belgium, in 1968, and the Ph.D. degree from Delft University of Technology, Delft, The Netherlands, in 1981.

In 1970 he became a Research Assistant and Lecturer at DUT, where he is now Associate Professor in the Department of Electrical Engineering. His current research interests are in VLSI and signal processing, particularly speech and image processing, filter design and modeling, systolic signal processors, and matrix equation solvers.

Rob J. Sluyter was born in Nijmegen, The Netherlands, on July 12, 1946. In 1968 he graduated in electronic engineering from the Eindhoven Instituut voor Hoger Beroepsonderwijs.

He joined Philips Research Laboratories, Eindhoven, The Netherlands, in 1962. Until 1978 he was a Research Assistant involved in data transmission and low bit rate speech coding. In 1978 he became a Staff Researcher engaged in speech analysis, synthesis, digital coding of speech, and digital signal processing. Since 1982 he has been engaged in research on medium bit rate coding of speech for mobile radio applications as a member of the Digital Signal Processing Group. His current interests are in digital signal processing for television signals.