[3] C. S. Burrus and T. W. Parks, "Time domain design of recursive digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 137-141, June 1970.
[4] F. N. Cornett, "First and second order techniques for the design of recursive filters," Ph.D. dissertation, Colorado State Univ., Ft. Collins, CO, 1976.
[5] D. C. Farden and L. L. Scharf, "Statistical design of nonrecursive digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 188-196, June 1974.
[6] E. Parzen, "Multiple time series Modeling," in *Multivariate Analysis II*, P. Krishnaiah, Ed. New York: Academic, pp. 389-409.
[7] W. C. Kellog, "Time domain design of nonrecursive least mean-square digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 155-158, June 1972.
[8] D. C. Farden and L. L. Scharf, "Authors reply to 'Comments on statistical design of nonrecursive digital filters,'" *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 495-496, Oct. 1975.
[9] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*. Berkeley, CA: Univ. California Press, 1958.
[10] K. Steiglitz, "Computer-aided design of recursive digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 123-129, June 1970.
[11] P. Thajchayapong and P. J. W. Rayner, "Recursive digital filter design by linear programming," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 107-112, Apr. 1973.
[12] L. R. Rabiner, N. Y. Graham, and H. D. Helms, "Linear programming design of IIR digital filters with arbitrary magnitude functions," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 117-123, Apr. 1974.
[13] D. E. Dudgeon, "Recursive filter design using differential correction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 443-448, Dec. 1974.
[14] H. Dubois and H. Leich, "On the approximation problem for recursive digital filters with arbitrary attenuation curve in the passband and the stopband," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 202-207, Apr. 1975.
[15] C. Charalambous, "Minimax optimization of recursive digital filters using recent minimax results," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 333-346, Aug. 1975.
[16] J. A. Cadzow, "Recursive digital filter synthesis via gradient based algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 349-356, Oct. 1976.
[17] H. Clergeot and L. L. Scharf, "Connections between classical and statistical methods of FIR digital filter design," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 463-465, Oct. 1978.
[18] K. Steiglitz and L. E. McBride, "A technique for identification of linear systems," *IEEE Trans. Automat. Contr.*, vol. AC-10, pp. 461-464, Oct. 1965.

Predictive Coding of Speech Signals and Subjective Error Criteria

BISHNU S. ATAL, SENIOR MEMBER, IEEE, AND MANFRED R. SCHROEDER, FELLOW, IEEE

Abstract—**Predictive coding methods attempt to minimize the rms error in the coded signal. However, the human ear does not perceive signal distortion on the basis of rms error, regardless of its spectral shape relative to the signal spectrum. In designing a coder for speech signals, it is necessary to consider the spectrum of the quantization noise and its relation to the speech spectrum. The theory of auditory masking suggests that noise in the formant regions would be partially or totally masked by the speech signal. Thus, a large part of the perceived noise in a coder comes from frequency regions where the signal level is low. In this paper, methods for reducing the subjective distortion in predictive coders for speech signals are described and evaluated. Improved speech quality is obtained: 1) by efficient removal of formant and pitch-related redundant structure of speech before quantizing, and 2) by effective masking of the quantizer noise by the speech signal.**

I. Introduction

FOR autocorrelated signals, such as speech, predictive coding [1]-[4] is an efficient method of encoding the signal into digital form. The coding efficiency is achieved by quantizing and transmitting only the signal which cannot be predicted from the already coded signal. In predictive coders, the power of the quantizer noise is proportional to the power of the prediction error. Thus, efficient prediction is important for minimizing the quantizer error. Small quantization error, however, does not ensure that the distortion in the speech signal is *perceptually* small; it is necessary to consider the *spectrum* of the quantization noise and its relation to the speech spectrum. The theory of auditory masking suggests that noise in the formant regions would be partially or totally masked by the speech signal. Thus, a large part of the perceived noise in a coder comes from the frequency regions where the signal level is low. Moreover, we can tolerate more distortion in the transitional segments in speech (where rapidly changing formants produce wider formant regions) in comparison to the steady segments.

In this paper, we discuss methods for modifying the spectrum of the quantization noise in a predictive coding system for speech to reduce the perceptible distortion introduced by such coders. The proper spectral shaping is realized by controlling the frequency response of the feedback network in the predictive coder independently of the predictor. The methods permit adaptive adjustment of the noise spectrum dependent

Manuscript received July 18, 1978; revised November 28, 1978.
B. S. Atal is with Bell Laboratories, Murray Hill, NJ 07974.
M. R. Schroeder is with the Drittes Physikalisches Institut, University of Göttingen, Göttingen, Germany, and Bell Laboratories, Murray Hill, NJ 07974.

0096-3518/79/0600-0247$00.75

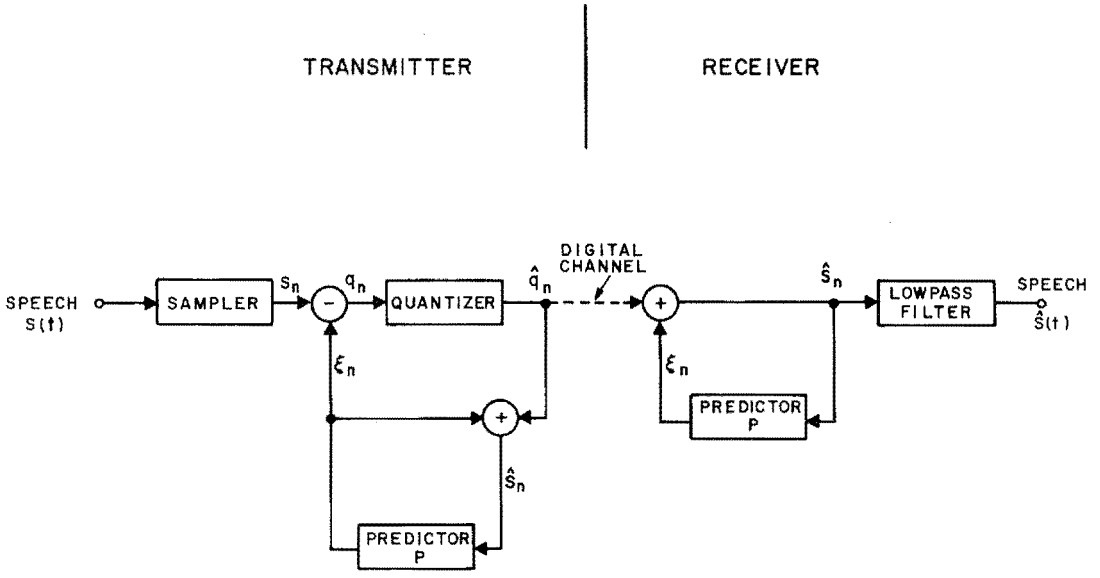

Fig. 1. Block diagram of a predictive coder.

on the time-varying speech spectrum. Improved speech quality is obtained both by exploiting auditory masking and by efficient prediction of formant and pitch-related redundancies in speech before quantizing.

II. Spectrum of Quantizing Noise in Predictive Coders

Fig. 1 shows a schematic diagram of a predictive coder. Its operation can be summarized as follows. The input speech signal is sampled to produce a sequence of sample values $s_0, s_1, \cdots, s_n, \cdots$. The predictor P forms a linear estimate of each sample value based on the previously decoded output sample values (reconstructed speech samples). This estimate ξ_n is subtracted from the actual sample value s_n and the resulting difference q_n is quantized and transmitted to the receiver. The quantized difference $\hat{q}_n$ is added back to the predicted value ξ_n both at the transmitter and at the receiver to form the next output sample $\hat{s}_n$. It is readily seen in Fig. 1 that, in the absence of channel errors, the coder output $\hat{s}_n$ is given by

$$\begin{aligned}\hat{s}_n &= \xi_n + \hat{q}_n \\ &= \xi_n + s_n - \xi_n + \delta_n \\ &= s_n + \delta_n \end{aligned} \tag{1}$$

where δ_n is the error introduced by the quantizer (difference between the output and the input of the quantizer) at the nth sample. Thus, the difference between the output and the input speech sample values is identical to the error introduced by the quantizer. Assuming that the spectrum of the quantizer error is white (a reasonable assumption, particularly if the prediction error is white), the noise at the output of the coder is also white.

III. Generalized Predictive Coder

The quantizer input q_n in Fig. 1 can be written as

$$\begin{aligned} q_n &= s_n - \sum_{k=1}^{m} \hat{s}_{n-k} a_k \\ &= s_n - \sum_{k=1}^{m} (s_{n-k} + \delta_{n-k}) a_k \\ &= s_n - \sum_{k=1}^{m} s_{n-k} a_k - \sum_{k=1}^{m} \delta_{n-k} a_k \end{aligned} \tag{2}$$

where the predictor P is represented by a transversal filter with m delays and m gains $a_1, a_2, \cdots, a_m$. The quantizer input thus consists of two parts: 1) the prediction error based on the prediction of the input s_n from its own past, and 2) the filtered error signal obtained by filtering of the quantizer error through the filter P. An easy way of modifying the spectrum of the quantizing noise is to use a filter F different from P for filtering the quantizer error [5], [6]. Fig. 2 shows such a generalized predictive coder. The quantizer input q_n is now given by

$$q_n = s_n - \sum_{k=1}^{m} s_{n-k} a_k - \sum_{k=1}^{m} \delta_{n-k} b_k \tag{3}$$

where $b_1, b_2, \cdots, b_{m'}$ are m' gain coefficients of the transversal filter F. We will assume that both $1 - F$ and $1 - P$ have their roots inside the unit circle. The coder output is now given as

$$\hat{s}_n = s_n + \sum_{k=1}^{m} (\hat{s}_{n-k} - s_{n-k}) a_k + \delta_n - \sum_{k=1}^{m} \delta_{n-k} b_k. \tag{4}$$

Representing Fourier transforms by upper case letters, (4) can be written in frequency-domain notations as

$$\hat{S} - S = \Delta \frac{1 - F}{1 - P}. \tag{5}$$

For $F = P$, the output noise is the same as the quantizer noise, and the two coders shown in Figs. 1 and 2 are identical. However, with $F \neq P$, the coder of Fig. 2 allows greater flexibility in controlling the spectrum of output noise based on the

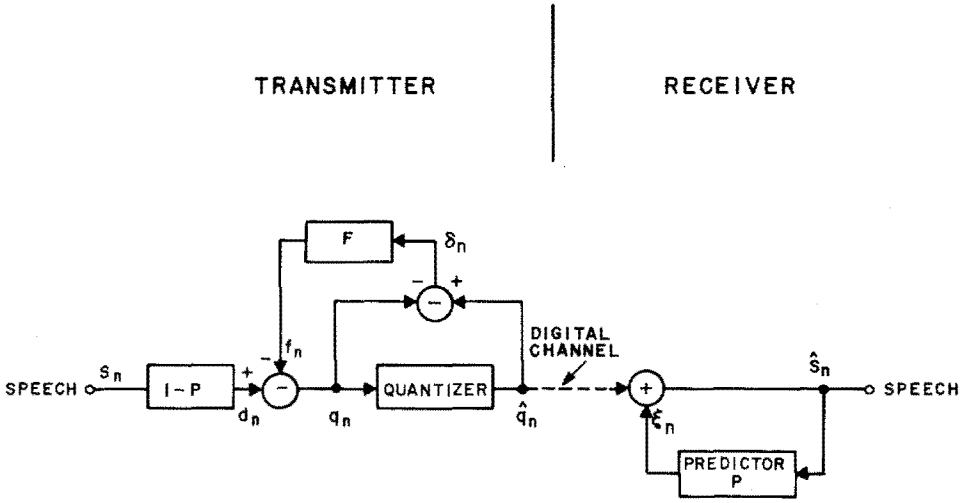

Fig. 2. Block diagram of a generalized predictive coder with adjustable noise spectrum.

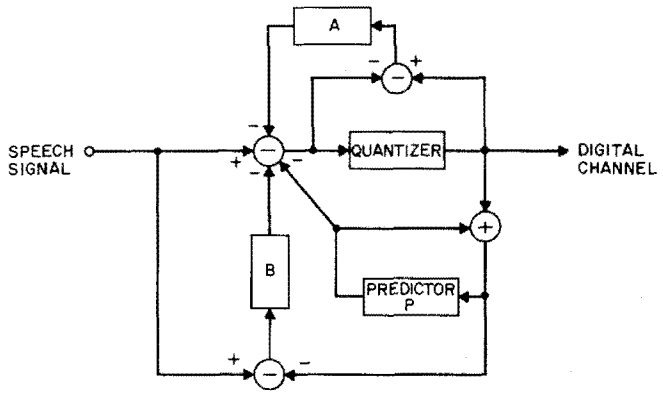

Fig. 3. Another configuration for the generalized predictive coder with adjustable noise spectrum.

choice of the feedback filter F.[1] Under the assumption that the quantizer noise is white, the spectrum of the coder output noise is determined only by the factor $(1 - F)/(1 - P)$ as shown in (5). Let the squared magnitude of this factor at a frequency f be $\Gamma(f)$. Then

$$\Gamma(f) = \left|[1 - F(e^{2\pi jfT})]/[1 - P(e^{2\pi jfT})]\right|^2 \tag{6}$$

where T is the sampling interval. Equation (6) implies an important constraint on the average value of $\log \Gamma(f)$, that is,

$$\frac{1}{f_s}\int_0^{f_s} \log \Gamma(f)\, df = 0 \tag{7}$$

where f_s is the sampling frequency. Expressed on a decibel scale, the average value of $\log \Gamma(f)$ is 0 dB. The proof of (7) is relatively straightforward [7] and is outlined below.

Consider the function $1 - F$ which is expressed in the z-transform notation as

$$1 - F(z) = 1 - \sum_{k=1}^{m'} b_k z^{-k} = \prod_{k=1}^{m'} (1 - z_k z^{-1}) \tag{8}$$

where z_k is the kth root of $1 - F(z)$. The function $\log [1 - F(z)]$ is given by

$$\log [1 - F(z)] = \sum_{k=1}^{m'} \log(1 - z_k z^{-1}). \tag{9}$$

Since $|z_k| < 1$ (all the zeros of $1 - F(z)$ are inside the unit circle), the right side of (9) can be expressed as a polynomial function of z^{-1}. Therefore,

$$\log [1 - F(z)] = \sum_{n=1}^{\infty} c_n z^{-n} = \sum_{n=1}^{\infty} c_n e^{-2\pi jfTn} \tag{10}$$

where $c_n = \Sigma_{k=1}^{m'} z_k^n$. The integral of $\log [1 - F(z)]$ over the frequency range from 0 to f_s is then given by

$$\int_0^{f_s} \log [1 - F(e^{2\pi jfT})]\, df = \sum_{n=1}^{\infty} c_n \int_0^{f_s} e^{-2\pi jfTn}\, df = 0. \tag{11}$$

Similarly, it can be shown that

[1] A somewhat different configuration of a predictive coder for controlling the spectrum of the output noise is shown in Fig. 3. It is easily verified that the Fourier transform of the output noise for this coder is given by $\Delta(1 - A)/(1 - B)$. In principle, the coder of Fig. 2 can be made equivalent to the coder of Fig. 3 by appropriate choice of the filter F. However, as a practical matter, one or the other coder may be simpler to implement depending on the choice for the spectrum of the output noise.

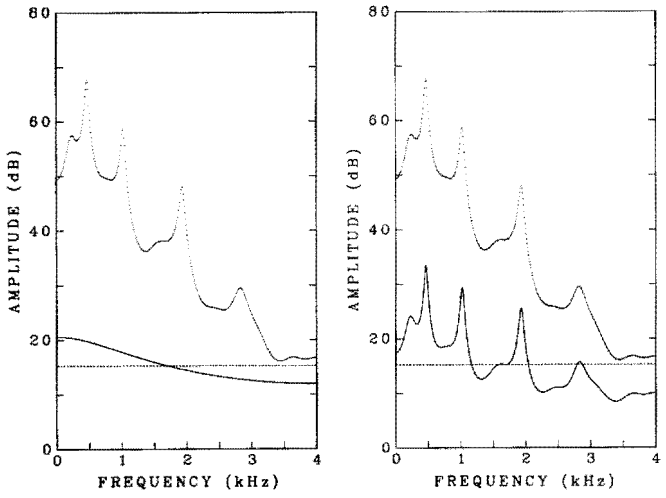

Fig. 4. Two possible shapes for the spectrum of output noise (solid curve) in the coder shown in Fig. 2. The average level of the logarithmic spectrum (shown as a dashed line) is the same in both cases. The speech spectrum is shown by the dotted curve.

$$\int_0^{f_s} \log\,[1 - P(e^{2\pi jfT})]\,df = 0. \tag{12}$$

Equation (7) follows directly from (11) and (12).

Assuming that the power of the quantizer noise δ_n is not changed significantly by the feedback loop—a desirable condition for satisfactory operation of the coder—the *average* value of *log* power spectrum of output noise is then determined solely by the quantizer and is not altered by the choice of the filter F or the predictor P. The filter F, however, redistributes the noise power from one frequency to another. Thus, reduction in quantizer noise at one frequency can be obtained only at the expense of increasing the quantizer noise at another frequency. Since a large part of perceived noise in a coder comes from the frequency regions where the signal level is low, the filter F can be used to reduce the noise in such regions while increasing the noise in the formant regions where the noise could be effectively masked by the speech signal. Some examples of the possible shapes for the spectrum of the quantizing noise together with the speech spectrum are illustrated in Fig. 4. In each case, the logarithmic spectrum of the quantizing noise has equal area above and below the average level shown by the dashed line.

IV. Application to Speech Signals

A. Selection of Predictor

Linear prediction is a well-known method of removing the redundancy in a signal. For speech, the prediction is done most conveniently in two separate stages [4], [8]: a first prediction based on the short-time spectral envelope of speech, and a second prediction based on the periodic nature of the spectral fine structure. The short-time spectral envelope of speech is determined by the frequency response of the vocal tract and for voiced speech also by the spectrum of the glottal pulse. The spectral fine structure arising from the quasi-periodic nature of voiced speech is determined mainly by the pitch period. The fine structure for unvoiced speech is random and cannot be used for prediction.

Prediction Based on Spectral Envelope

Prediction based on the spectral envelope involves relatively short delays. The predictor can be characterized in the z-transform notation as

$$P_s(z) = \sum_{k=1}^{p} a_k z^{-k} \tag{13}$$

where z^{-1} represents a delay of one sample interval and $a_1, a_2, \cdots, a_p$ are p predictor coefficients. The value of p typically is 10 for speech sampled at 8 kHz. A higher value may often be desirable.

The input to the quantizer in Fig. 2 consists of two parts: 1) the prediction error d_n based on the prediction of the input s_n from its own past, and 2) the filtered error signal f_n obtained by filtering of the quantizer noise δ_n through the filter F. Under the assumption that the quantizer noise is uncorrelated with the prediction error, the total power ϵ_q at the input to the quantizer is the sum of the powers in the prediction error d_n and the filtered noise f_n. That is,

$$\epsilon_q = \epsilon_p + \epsilon_f, \tag{14}$$

where ϵ_p is the power in the prediction error and ϵ_f is the power in the filtered noise. It is our experience that, for satisfactory operation of the coder, ϵ_f should be less than ϵ_p. The power in the filtered noise is determined both by the power in the quantizer error δ_n and the power gain G of the filter F. The power gain G equals the sum of the squares of the filter coefficients. For $F = P_s$, the power gain is usually large and can often exceed 200. Such a high power gain causes excessive feedback of the noise power to the quantizer input, particularly for coarse quantizers, resulting in poor performance of the coder. Excessive feedback can be prevented by requiring that the power gain of the filter $P_s(z)$ is not large. The reason for the high power gain is as follows.

The spectrum of the filter $1 - P_s(z)$ is approximately the reciprocal of the speech spectrum (within a scaling constant). The low-pass filter used in the analog-to-digital conversion of the speech signal forces the reciprocal spectrum (and thus $|1 - P_s(z)|$) to assume a high value in the vicinity of the cutoff frequency of the filter. The power gain, which is equal to the integral of the power spectrum $|1 - P_s(e^{2\pi jfT})|^2$ with respect to the frequency variable f, thus also becomes large.

These artificially high power gains will not arise if the low-pass filter used in the sampling process was an ideal low-pass filter with a cutoff frequency exactly equal to the half the sampling frequency. The amplitude-versus-frequency response of a practical low-pass filter falls off gradually. The computed covariance matrix used in LPC analysis therefore has missing components corresponding to the speech signal rejected by the low-pass filter. The missing high-frequency components produce artificially low eigenvalues of the covariance matrix corresponding to eigenvectors related to such components. The high power gain of P_s is precisely caused by the small eigenvalues. The covariance matrix of the low-pass filtered speech is nearly singular thereby resulting in a nonunique solution of the predictor coefficients. Thus, a variety of different predictor coefficients can approximate the speech

spectrum equally well in the passband of the low-pass filter. We wish to avoid solutions which lead to high power gains of the predictor P_s.

The ill-conditioning of the covariance matrix can be avoided by adding to the covariance matrix another matrix proportional to the covariance matrix of high-pass filtered white noise. We define a new covariance matrix $\hat{\Phi}$ (with its (ij) term represented by $\hat{\phi}_{ij}$) and a new correlation vector $\hat{c}$ (with its ith term represented by $\hat{c}_i$) by the equations

$$\hat{\phi}_{ij} = \phi_{ij} + \lambda \epsilon_{\min} \mu_{i-j} \tag{15}$$

and

$$\hat{c}_i = c_i + \lambda \epsilon_{\min} \mu_i \tag{16}$$

where

$$\phi_{ij} = \langle s_{n-i} s_{n-j} \rangle,$$

$$c_i = \langle s_n s_{n-i} \rangle,$$

λ is a small constant (suitable values are in the range 0.01–0.10), $\epsilon_{\min}$ is the minimum value of the mean-squared prediction error, μ_i is the autocorrelation of the high-pass filtered white noise at a delay of i samples, and $\langle\ \rangle$ indicates averaging over the speech samples contained in the analysis segment. Ideally, the high-pass filter should be the filter complimentary to the low-pass filter used in the sampling process. We have obtained reasonably satisfactory results with the high-pass filter $[\frac{1}{2}(1 - z^{-1})]^2$. For this filter, the autocorrelations are $\mu_0 = \frac{3}{8}$, $\mu_1 = -\frac{1}{4}$, $\mu_2 = \frac{1}{16}$, and $\mu_k = 0$ for $k > 2$. By making the scale factor on the noise covariance matrix in (15) and (16) proportional to the mean-squared prediction error, we find that it is possible to use a fixed value of λ. The results are not very sensitive to small variations in the value of λ. The minimum value of the mean-squared prediction error is determined by the Cholesky decomposition [8] of the original covariance matrix $[\phi_{ij}]$. A modified form [9] of the covariance method is used to determine the predictor coefficients from the new covariance matrix $\hat{\Phi}$. The first two steps in this modifed procedure are identical to the usual covariance method [8]. That is, the matrix $\hat{\Phi}$ is expressed as a product of a lower triangular matrix L and its transpose L^t by Cholesky decomposition and a set of linear equations $Lq = \hat{c}$ is solved. The partial correlation at a delay m is obtained from

$$r_m = \frac{q_m}{\left[\langle s_n^2 \rangle - \sum_{j=1}^{m-1} q_j^2\right]^{1/2}} \tag{17}$$

where q_m is the mth component of q. The partial correlations are transformed to predictor coefficients using the well-known relation between the partial correlations and the predictor coefficients for all-pole filters [7, p. 110]. The modified procedure ensures that all of the zeros of the polynomial $1 - P_s(z)$ are inside the unit circle. Using the above procedure, reasonable power gains are realized without introducing significant bias in the spectrum at lower frequencies.[2] Examples of spectral envelopes of speech computed for $\lambda = 0$ (uncorrected) and for $\lambda = 0.05$ with the high-pass filter $[\frac{1}{2}(1 - z^{-1})]^2$ are illustrated in Fig. 5. The power gain of P_s decreased from 204.6 for $\lambda = 0$ to 12.6 for $\lambda = 0.05$. The speech signal was sampled at a rate of 10 kHz. The anti-aliasing filter had attenuation of 3 dB at 4.2 kHz and more than 40 dB at 5 kHz.

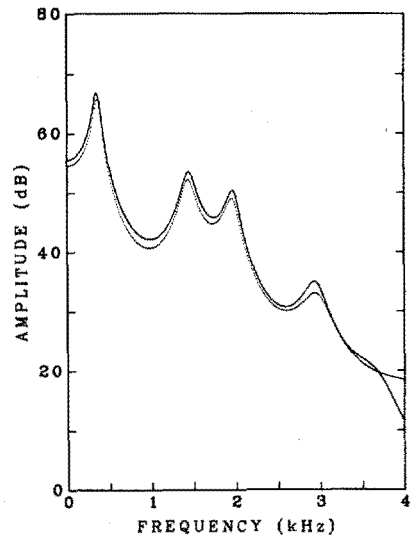

Fig. 5. Spectral envelopes of speech based on LPC analysis with high-frequency correction for $\lambda = 0$ (solid curve) and for $\lambda = 0.05$ (dotted curve). The power gains in the two cases are 204.6 and 12.6, respectively.

Prediction Based on Spectral Fine Structure

Adjacent pitch periods in voiced speech show considerable similarity. The quasi-periodic nature of the signal is present—although to a lesser extent—in the difference signal obtained after prediction on the basis of spectral envelope. The periodicity of the difference signal can be removed by further prediction. The predictor for the difference signal can be characterized in the z-transform notation by

$$P_d(z) = \beta_1 z^{-M+1} + \beta_2 z^{-M} + \beta_3 z^{-M-1} \tag{18}$$

where M represents a relatively long delay in the range 2 to 20 ms. In most cases, this delay would correspond to a pitch period (or possibly, an integral number of pitch periods). The degree of periodicity in the difference signal varies with frequency. The three amplitude coefficients β_1, β_2, and β_3 provide a frequency-dependent gain factor in the pitch-prediction process. We found it necessary to use at least a third-order predictor for pitch prediction. The difference signal after prediction based on spectral envelope has a nearly flat spectrum up to half the sampling frequency. Due to a fixed sampling frequency unrelated to pitch period, the individual samples of the difference signal do not show a high period-to period correlation. The third-order pitch predictor provides an interpolated value with a much higher correlation than the individual samples. Higher order pitch predictors provide even better improvement in the prediction gain.

Let the nth sample of the difference signal after the first ("formant") prediction be given by

[2]Another possible solution, namely, undersampling of the speech signal, for avoiding excessive power gains was suggested by one of the reviewers of this paper. We did not try this solution and are unable to comment on its effectiveness.

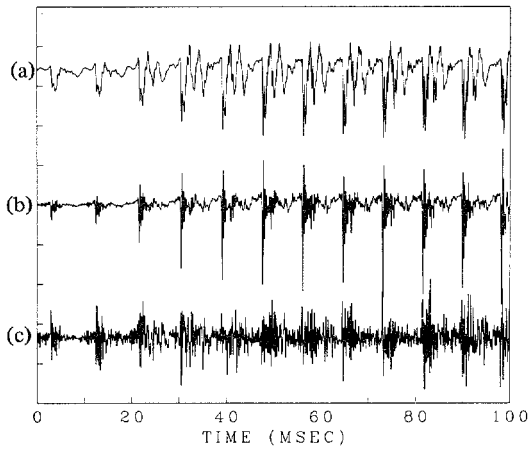

Fig. 6. (a) Speech waveform. (b) Difference signal after prediction based on spectral envelope (amplified 10 dB relative to the speech waveform). (c) Difference signal after prediction based on pitch periodicity (amplitied 20 dB relative to the speech waveform).

$$d_n = s_n - \sum_{k=1}^{p} a_k s_{n-k}. \tag{19}$$

The delay M of the predictor $P_d(z)$ is defined as the delay for which the normalized correlation coefficient between d_n and d_{n-M} is highest. The coefficients β_1, β_2, and β_3 are determined by minimizing the mean-squared prediction error between d_n and its predicted value. The minimization procedure leads to a set of simultaneous *linear* equations in the three unknowns β_1, β_2, and β_3.

Examples of prediction error signals after each stage of prediction together with the original speech signal are illustrated in Fig. 6. The prediction error after the first prediction based on the spectral envelope is amplified by 10 dB in the display. The prediction error after the second prediction based on pitch periodicity is amplified by an additional 10 dB. The prediction error after pitch prediction is quite noise-like in nature. Its spectrum—including both envelope and fine structure—is nearly white.

B. Selection of the Feedback Filter

The filter F providing feedback of the quantizer error in Fig. 2 can be chosen to minimize an error measure in which the noise is weighted according to some subjectively meaningful criterion. For example, an effective noise-to-signal ratio can be defined by weighting the noise power at each frequency f by a function $W(f)$. For a fixed quantizer, the spectrum of the output noise is proportional to $|(1 - F)/(1 - P)|^2$. The signal spectrum is proportional to $|1/(1 - P)|^2$. The ratio of noise power to signal power at any frequency f is proportional to

$$G(f) = |1 - F(e^{2\pi jfT})|^2. \tag{20}$$

One could choose F to minimize

$$E_n = \int_0^{f_s} G(f)\, W(f)\, df \tag{21}$$

under the constraint

$$\int_0^{f_s} \log G(f)\, df = 0. \tag{22}$$

The minimum is achieved if

$$\log G(f) = -\log W(f) + \frac{1}{f_s} \int_0^{f_s} \log W(f)\, df. \tag{23}$$

The function $1 - F$ is the minimum-phase transfer function with spectrum $G(f)$ and can be obtained by direct Fourier transformation or spectral factorization. A particularly simple solution to this problem is obtained by transforming $1/G(f)$ to an autocorrelation function by Fourier transformation. By using a procedure similar to LPC analysis, the autocorrelation function can be used to determine a set of predictor coefficients. The predictor coefficients determined in this manner are indeed the desired filter coefficients b_k for the filter F. The solution is considerably simplified also if the noise-weighting function $W(f)$ is expressed in terms of a filter transfer function whose poles and zeros lie inside the unit circle.

C. Generalized Predictive Coder for Speech

The block diagram of a generalized predictive coder for speech signals is shown in Fig. 7. The high frequencies in the speech signal are preemphasized with a filter $1 - 0.4z^{-1}$ prior to encoding. At the output, the high frequencies are deemphasized back with an inverse filter $(1 - 0.4z^{-1})^{-1}$. The speech samples are filtered by the filter $1 - P_s$ to produce the difference signal after the first prediction based on the short-time spectral envelope of the speech signal. The second predictor P_d based on pitch periodicity forms an estimate d_n' of the difference signal at the nth sample based on the past quantizer output samples. The quantizer error is filtered and peak-limited to produce the sample f_n. The composite signal $q_n = d_n - d_n' - f_n$ is quantized by an adaptive three-level quantizer. The parameters of the quantizer are selected to be optimum [10] for a Gaussian input with standard deviation equal to the rms value of the prediction error (after prediction by both P_s and P_d). For a uniformly spaced three-level quantizer, the optimum spacing between the quantizer output levels is 1.22 times the rms value of the prediction error. Both predictors and the quantizer parameters are reset once every 10 ms. An interval of 20 ms is used to minimize the prediction error in determining the predictor parameters. For a total of 40 s of speech spoken by two male and two female speakers, the average prediction gains for the two predictors—based on spectral envelope and pitch periodicity—were found to be 11.0 and 5.6 dB, respectively. The average signal-to-noise ratio (SNR) of the three-level quantizer was found to be 6.1 dB. The speech was relatively noise-free with an average SNR of approximately 40 dB.

It can be shown that the stability of the feedback loop is ensured only if the function $1 - F$ is a minimum-phase transfer function [11]. Several methods of ensuring the stability of the feedback loop were considered for situations where the minimum-phase requirement could not be satisfied. A simple and effective solution was found to be to limit the peak

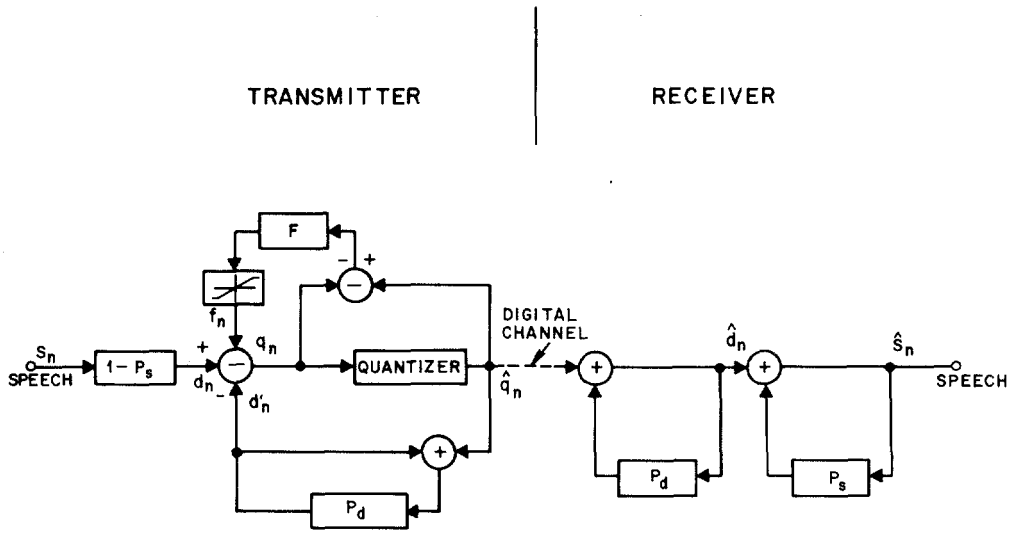

Fig. 7. Block diagram of a generalized predictive coder for speech.

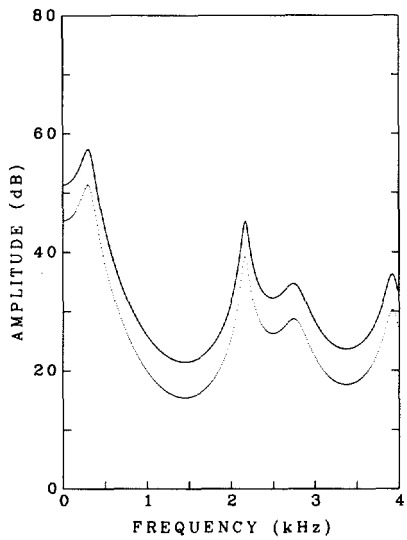

Fig. 8. Spectral envelopes of output quantizing noise (dotted curve) and the corresponding speech spectrum (solid curve) for $F = 0$.

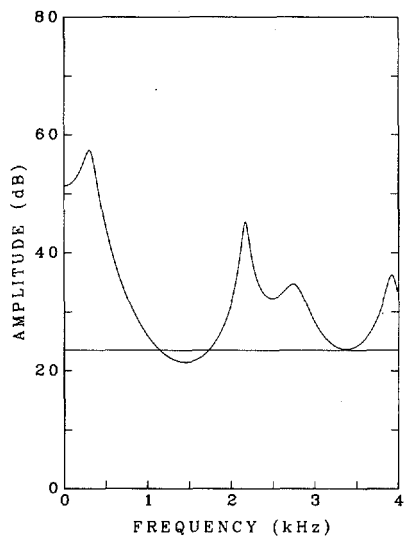

Fig. 9. Spectral envelope of the speech signal and flat quantization noise ($F = P_s$).

amplitude of the samples f_n. The peak limiter in the feedback loop limits the samples f_n to a maximum value of twice the rms value of the prediction error. For one of the choices for the filter F, several instances of instability in the feedback loop were encountered without the peak limiter. However, our results based on extensive testing with the speech data revealed no significant increase of the quantization noise after inclusion of the peak limiter in the feedback loop.

Several choices for the feedback filter F in Fig. 7 were investigated. Some of the more interesting ones are discussed below.

1) Assume that the noise-weighting function $W(f)$ in (23) is constant for all frequencies. This would be a good choice if our ears were equally sensitive to quantizing distortion at all frequencies. The solution is $F = 0$. In this case, the coder output noise has the same spectral envelope as the original speech (but at a lower level) as shown in Fig. 8. This particular choice leads to a signal-to-noise ratio of 13 dB with a three-level quantizer. Subjectively, the reconstructed speech is quite noisy. But the distortion is heard mostly at low frequencies implying that $W(f) = constant$ is a good choice at the higher frequencies.

2) Assume that the distortion is subjectively more important in the formant regions and less important in the regions between formants. As an extreme case, let $W = |1 - P_s|^{-2}$ which leads to $F = P_s$—the same system as the original predictive coder shown in Fig. 1. For a given quantizer, such a choice results in minimum unweighted noise power in the reconstructed speech. For the three-level quantizer, an average SNR of 23 dB is obtained. Subjectively, the reconstructed speech from the coder is much less noisy than for the case $F = 0$. But, the quantization noises in the two cases sound very different. For $F = P_s$, the spectrum of the output noise (see Fig. 9) is white, producing a very high SNR at the formants but a poor one in between the formants.

3) Select F somewhere in between the two extreme choices discussed above. As an example, let

$$F = P_s(\alpha z^{-1}) \tag{24}$$

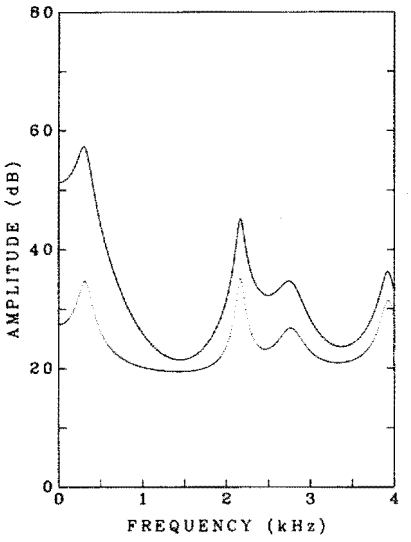

Fig. 10. An example showing the envelope of the output noise spectrum (dotted curve) shaped to reduce perceived distortion [F as in (24)] and the corresponding speech spectrum (solid curve).

where α is an additional parameter introduced to increase the bandwidths of the zeros of $1 - F$. An increase of 400 Hz in bandwidth was found to be satisfactory. The feedback filter F changes from $F = P_s$ for $\alpha = 1$, to $F = 0$ for $\alpha = 0$. An example of the envelope of the output noise spectrum together with the corresponding speech spectrum is shown in Fig. 10. In comparison to the previous case ($F = P_s$), the SNR is improved in the frequency regions where the speech spectral level is low at the expense of decreased SNR in the formant regions. The choice of F shown in (24) results in an average SNR of approximately 21 dB–lower than for $F = P_s$ but higher than for $F = 0$. Subjectively, the quality of the reconstructed speech in this case is much better than for either of the other two cases. Comparisons with PCM-encoded speech show that the speech quality with the three-level quantizer is close to that of log-PCM encoded speech at 7 bits/sample with a SNR of 33 dB. The nonuniform noise spectrum together with the predictive coding thus produces a gain of about 12 dB in the subjective SNR.

The above examples were meant to illustrate the potential benefits of controlling the noise spectrum in predictive coders. The three choices of F discussed here are somewhat arbitrary. The optimum choice of F should be determined from measurements reflecting perceptual masking of quantizing noise by speech at different frequencies and signal levels.

V. Concluding Remarks

Predictive coding is a promising approach for speech coding. It provides a practical method of realizing substantial savings in transmission requirements of digitized speech without compromising the speech quality while, at the same time, providing robust performance across different speakers and spoken material. This paper focuses on a hitherto neglected aspect of predictive coding systems–namely, the subjectively weighted spectrum of the noise introduced by the coder. A large part of the perceived coder noise comes from the frequency regions where the noise is not masked by the speech signal. A simple modification of the conventional predictive coder permits adjustment of the noise spectrum. It is shown that considerable improvement in speech quality can be achieved by proper shaping of the noise spectrum dependent on the short-time spectral envelope of the speech signal.

Efficient prediction of the redundant structure in a signal is important for proper functioning of a predictive coder. This paper describes methods for linear prediction of speech based on the nonuniform nature of its spectral envelope and on its quasi-periodic nature. The resulting prediction error is only 1/100 in power as compared to the speech signal. Low prediction error combined with proper shaping of the output noise spectrum produces speech with a three-level predictive quantizer which is approximately equivalent in quality to 7-bit log-PCM encoded speech.

References

[1] P. Elias, "Predictive coding," *IRE Trans. Inform. Theory*, vol. IT-1, pp. 16–33, Mar. 1955.
[2] C. C. Cutler, "Differential quantization of communication," U.S. Patent 2 605 361, 1952.
[3] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals," in *Proc. 1967 Conf. Speech Commun. and Processing*, Nov. 1967, pp. 360–361.
[4] —, "Adaptive predictive coding of speech signals," *Bell Syst. Tech. J.*, vol. 49, pp. 1973–1986, 1970.
[5] C. C. Cutler, "Transmission systems employing quantization," U.S. Patent 2 927 962, 1960.
[6] E. G. Kimme and F. F. Kuo, "Synthesis of optimum filters for a feedback quantization system," *IEEE Trans. Circuit Theory*, vol. CT-10, pp. 405–413, Sept. 1963.
[7] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer-Verlag, 1976, p. 130.
[8] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, Aug. 1971.
[9] B. S. Atal, "On determining partial correlation coefficients by the covariance method of linear prediction," *J. Acoust. Soc. Amer.*, vol. 62, suppl. 1, paper BB12, p. S64, Fall 1977.
[10] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inform. Theory*, vol. IT-6, pp. 7–12, Mar. 1960.
[11] D. Mitra, personal communication.