THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 10, December 1975
Printed in U.S.A.

An Efficient Linear-Prediction Vocoder

By M. R. SAMBUR

(Manuscript received April 29, 1975)

A primary interest in any method for producing synthetic speech is to minimize the number of bits per second required to generate acceptable quality speech. An efficient method for transmitting the linear-prediction parameters has been found by using the techniques of differential PCM. *Using this technique, speech transmission is achieved employing fewer than 1500 bits/s. Further reductions in the linear-prediction storage requirements can be realized at a cost of higher system complexity by transmission of the most significant eigenvectors of the parameters. This technique in combination with differential* PCM *can lower the storage to 1000 bits/s.*

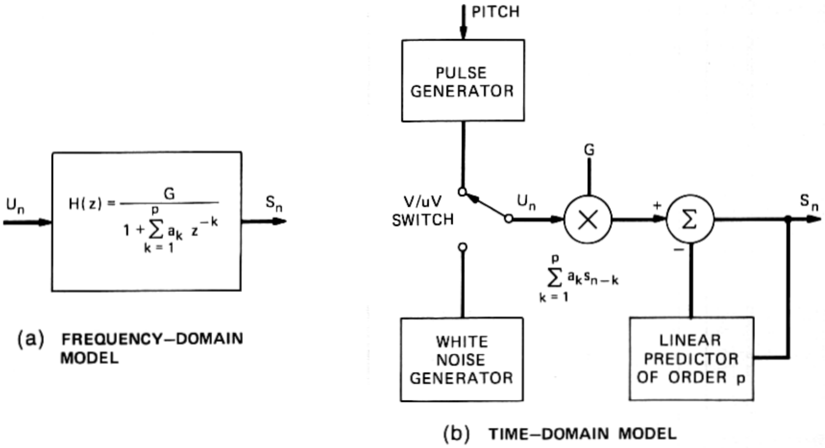
I. INTRODUCTION

The method of linear prediction has proved quite popular and successful for use in speech compression systems.[1–4] In this method, speech is modeled as the output of an all-pole filter $H(z)$ that is excited by a sequence of pulses separated by the pitch period for voiced sounds, or pseudo-random noise for unvoiced sounds. These assumptions imply that within a frame of speech the output speech sequence is given by

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + u_n,$$

where p is the number of modeled poles, u_n is the appropriate input excitation, and the a_k's are the coefficients characterizing the filter (linear prediction coefficients). Figure 1 illustrates the frequency-domain, as well as the equivalent time-domain, model of linear-prediction speech production. To account for the nonstationary character of the speech waveform, the parameters a_k of the modeled filter are periodically updated during successive speech frames.* Generation of speech in this method requires a knowledge of the pitch, the filter

* A frame is a segment of speech thought adequate to assume stationarity of the speech process. Typical frame lengths employed range from 10 to 30 ms.


PITCH
PULSE GENERATOR
G
U_n
$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}}$
S_n
V/uV SWITCH
U_n
+
−
$\sum_{k=1}^{p} a_k s_{n-k}$
S_n
WHITE NOISE GENERATOR
LINEAR PREDICTOR OF ORDER p

(a) FREQUENCY–DOMAIN MODEL

(b) TIME–DOMAIN MODEL

Fig. 1—Discrete model of speech production as employed in linear prediction.

parameters, and the gain of the filter (amplitude of input excitation) in each speech frame.

A primary interest in any method for producing synthetic speech is to minimize the number of bits per second needed to generate acceptable quality speech. The smaller the information storage requirements (bits per second), the more attractive the system becomes for the important applications of voice answer-back and speech transmission.[5] To achieve the minimum storage requirement for a given system, an efficient means of quantizing the generating parameters must be determined. Using conventional pulse code modulation (PCM) techniques in which the amplitude of each parameter is uniformly quantized into 2^B levels, it has been found necessary to allot at least five bits ($B = 5$) of information for both pitch and gain and at least 11 bits for each a_k.[1] The corresponding storage requirements for this method of quantization of the linear-prediction (LPC) parameters is unacceptable for many applications, and an improved scheme for quantizing the parameters is needed.

For the usual 12-pole linear-prediction representation, the dominant portion of storage is allotted to the filter coefficients (132 bits per frame in the PCM method of information transmission). The extremely fine quantization of the a_k's is necessary because small perturbations in the coefficients can sometimes cause radical changes in the important pole frequencies of the modeled filter $H(z)$ and may even cause the filter to become unstable (poles outside the unit circle). To overcome the limitations of quantizing the predictor (filter) coefficients, it has been found quite profitable to transmit different but informationally equivalent sets of parameters.[4,6] The most suitable parameters have been experimentally determined to be the log-area ratio coefficients g_i.[4] These coefficients are nonlinearly related to the filter coefficients by

$$g_i = \log \frac{1 + k_i}{1 - k_i}, \tag{1}$$

where the k_i's are termed the parcor coefficients.[1,2,4,6] If we denote $a_i^{(j)}$ as the ith linear prediction coefficient for a jth-pole linear-prediction model, then

$$k_i = a_i^{(i)}. \tag{2}$$

The parcor coefficients have the very important property that if

$$|k_i| < 1, \qquad i = 1, \cdots, p, \tag{3}$$

then it is guaranteed that the linear-prediction filter is stable.[4] Thus, small perturbations in the parcor coefficients or log-area coefficients will not affect the stability of the modeled filter.

Since the log-area coefficients are nonlinearly related to the filter coefficients, transmission of the log-area parameters is equivalent to a nonuniform quantization of the linear-prediction coefficients. By transmitting the log-area parameters, the number of bits allotted to the filter parameters can be effectively reduced by nearly $\frac{1}{2}$.[3,4,6] In this paper, we offer two additional methods of quantization of the necessary synthesis parameters (pitch, gain, and filter coefficients) that can even further reduce the storage requirements of a linear-prediction vocoder. One proposed method of quantization uses the technique of differential PCM (DPCM) to transmit the linear-prediction parameters. This scheme exploits the fact that the LPC parameters are themselves predictable from previous samples. Using this method, speech transmission that is practically equivalent to the unquantized synthesis can be achieved using fewer than 2000 bits/s.

The second method of transmission exploits the redundancy between the linear-prediction parameters. The LPC parameters can be predicted not only from the given parameter's past values, but also in some sense from values of the other parameters. The suggested scheme involves the transmission (using DPCM techniques) of the most significant eigenvectors of the log-area parameters. For the typical speech utterance, the space of the 12 log-area coefficients can be effectively represented by only the first five or six eigenvectors. The transmission requirement for this method is fewer than 1200 bits/s.

The organization of this paper is as follows. In Section II, we briefly discuss the concept of DPCM coding. In Section III, we show that DPCM coding offers a significant improvement over PCM coding for transmission of the linear-prediction parameters. In Section IV, the results are presented of a synthetic speech experiment using the proposed DPCM scheme. In Section V, we discuss several methods of DPCM coding that are more suitable for real-time implementation. Included in this section is a discussion of adaptive quantization (ADPCM) and adaptive DPCM prediction. In Section VI, we discuss the method of orthogonal linear prediction. The results of synthetic speech experiments are included in this section. Finally, we conclude with a summary and discussion of the results presented in the paper.

II. DIFFERENTIAL PULSE CODE MODULATION

The idea of differential PCM is similar in philosophy to the concept employed in linear-prediction speech analysis. In DPCM, we assume that the transmitted parameter in a given frame of interest can be estimated by a linear combination of the parameter in previous frames.[7] If we let x_r denote the value of the transmission parameter x in the rth frame (where x can represent pitch, gain, log-area coefficients, or whatever), then this assumption implies

$$x_r \approx \hat{x}_r = \sum_{i=1}^{n} b_i x_{r-i}, \tag{4}$$

where n is the order of the DPCM prediction analysis. The DPCM technique calls for the transmission of the difference between the predicted value $\hat{x}_r$ and the true value x_r.

Figure 2 illustrates the structure of the DPCM coding system. In the implementation of a DPCM scheme, a feedback path around the quantizer is used to ensure that the error in the reconstructed (quantized) signal $\tilde{x}_r$ is precisely the quantization error for the difference signal $e_r = x_r - \hat{\tilde{x}}_r$, where $\hat{\tilde{x}}$ is the predicted value based upon the quantized signal $\tilde{x}_r$. The predictor coefficients b_i are chosen to minimize the power of the difference signal e_r. The mathematical analysis required for the solution of the optimum set of b_i's is exactly the same as the analysis for the calculation of the linear-prediction coefficients, a_i, $i = 1, \cdots, p$. The determination of the b_i's is made by solving the familiar correlation equations:

$$\sum_{i=1}^{n} b_i \sum_{r=n}^{N} x_{r-i} x_{r-k} = - \sum_{r=n}^{N} x_r x_{r-k}, \qquad 1 \leqq k \leqq n, \tag{5}$$

where N is the number of frames in the utterance.

The advantage of DPCM coding is obvious when one realizes that, if x_r can be accurately estimated from previous samples, the information necessary for transmission (as expressed by the difference signal $x_r - \hat{\tilde{x}}$) is necessarily less than the information required for coding the signal without exploiting its predictability. The advantage of

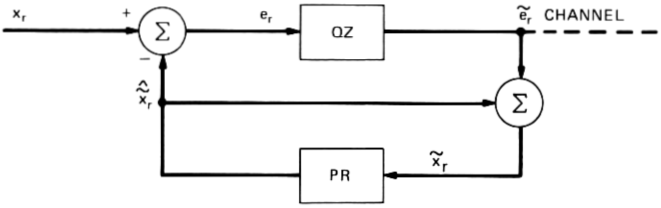


Fig. 2—Differential PCM (QZ = quantizer; PR = predictor; = $\hat{\hat{x}}_r = \sum_{j=1}^{r} b_j x_{r-j}$).

DPCM coding can be precisely specified by noting that, for a given fineness of quantization, the quantization error is proportional to the variance of the signal present at the quantizer.[7] Thus, the improvement in performance (as measured by the frequently used standard of signal-to-quantization-error ratio) using DPCM strategy over straight PCM coding is given by the ratio of the variance (power) of x_r to that of the difference signal

$$G = \frac{\langle x_r^2 \rangle}{\langle (x_r - \hat{\hat{x}})^2 \rangle}. \tag{6}$$

Using the optimum predictors b_i, the resulting gain over PCM is approximately* given by

$$G_{\text{opt}} = \left(1 - \sum_{k=1}^{n} \frac{b_i C_i}{C_0}\right)^{-1} = \frac{\langle x_r^2 \rangle}{\langle (x_r - \hat{x}_r)^2 \rangle}, \tag{7}$$

where

$$C_i = \sum_{r=n}^{N} x_r x_{r-i}. \tag{8}$$

For equal standards of synthetic speech quality, the gain obtained by using a DPCM strategy over that of PCM coding can be traded off to reduce the rate of information transmission. Of course, for $G < 1$, it is disadvantageous to use DPCM coding. However, for the transmission of parameters that are reasonably smooth in their variation from one transmission frame to the next, it is guaranteed that DPCM coding is superior to PCM coding. In the next section, we demonstrate the efficiency of DPCM techniques for the coding of the linear-prediction speech parameters.

III. DPCM IMPROVEMENT IN CODING LPC PARAMETER

To illustrate the efficiency of DPCM techniques in the coding of the synthesis parameters, Fig. 3 shows the improvement factor G_{opt} in decibels as a function of the number of DPCM predictors. The figure

* The gain is approximate because the effects of the quantizer in Fig. 2 are ignored.

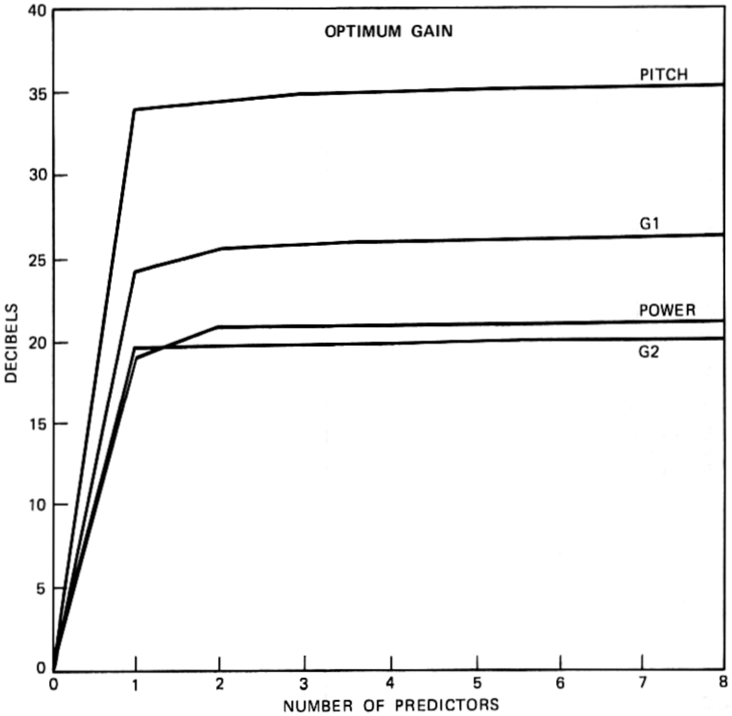


Fig. 3—G_{opt} for the sentence, "May we all learn a yellow lion roar."

shows G_{opt} for the first two log-area coefficients (g_1 and g_2),[*] pitch period and power[†] for the all-voiced utterance, "May we all learn a yellow lion roar." The improvement factor was calculated by considering each particular parameter across the entire sentence and then calculating the optimum predictors using eq. (5) and G_{opt} using eq. (7). The results depicted in Fig. 3 are for a male speaker, but the results are typical of those obtained for other male and female speakers. For the complete ensemble of parameters necessary to produce synthetic speech (12 log-area coefficients, pitch, and power),[‡] the set of improvement factors were all significantly greater than 1.

Figure 4 shows a typical plot of the improvement factor calculated for a sentence containing unvoiced sounds, "Few thieves are never

[*] The parameters were calculated at the rate of 50 samples per second. The filter parameter was calculated by the covariance method (Ref. 1), and pitch was measured by a method developed by B. S. Atal (Ref. 8).

[†] Power is defined as the energy in the speech frame. For the synthetic system employed, it is more convenient to transmit power instead of the amplitude of the input excitation.

[‡] Log-area coefficients were transmitted because of their optimum quantization properties.

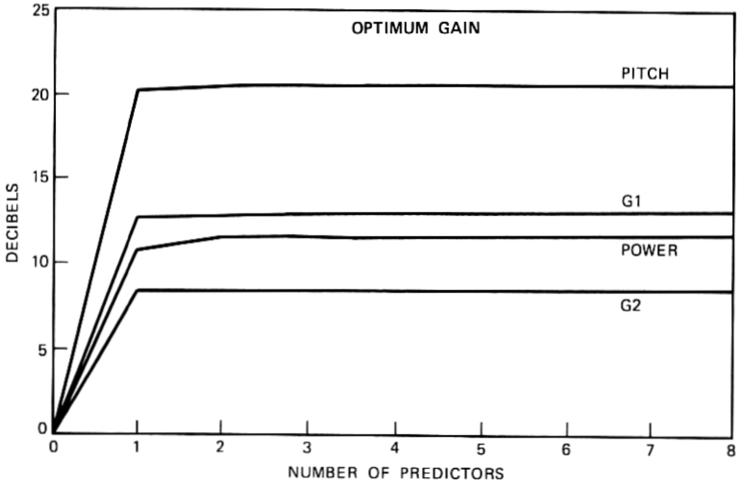


Fig. 4—G_{opt} for the sentence, "Few thieves are never sent to the jug."

sent to the jug." In this sentence, the DPCM improvement over PCM coding is not as dramatic as for the all-voiced sentence. The reason for the decreased values of G_{opt} is that the synthesis parameters tend to change sharply during the unvoiced-voiced transition. Thus, during the transition region there is an abrupt reduction in the correlation between successive samples, and very little information can be gained about the signal from past values. Another reason for the reduced values of G_{opt} is that the variation of the LPC parameters during un-

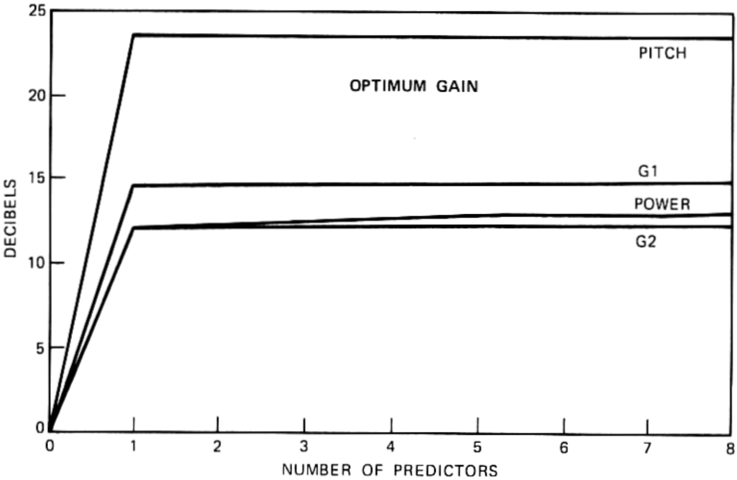


Fig. 5—G_{opt} for the sentence, "Few thieves are never sent to the jug." A separate DPCM analysis is used in each unvoiced and voiced segment.

voiced sounds is inherently more random than during voiced sounds and is thus less predictable. Fortunately, during unvoiced regions the quality of the synthesized speech is more tolerant to quantization noise than during voiced regions.[4] Thus, the diminished values of the G_{opt}'s is not as significant as might at first appear.

One method of increasing the improvement factor for utterances containing unvoiced sounds is to update the DPCM predictors whenever the spectral properties of the speech signal change from unvoiced to voiced sounds. Figure 5 shows G_{opt} for the same sentences as were used to obtain the results of Fig. 4, but in this figure the optimum DPCM predictors were separately calculated for each different section of unvoiced and voiced speech. The improvement factor for this form of DPCM coding is about 5 dB better than a single calculation of the predictors. In a later section of the paper, we discuss another method for updating or adapting the DPCM predictors to the changing spectral properties of the speech signal.

IV. SPEECH SYNTHESIS

4.1 *Synthesizer*

The improvement factors for the LPC parameters demonstrate that DPCM coding is superior to PCM coding. To confirm the results of the G_{opt} experiments, a synthetic speech system was constructed in the manner illustrated in Fig. 2. To take advantage of the fact that the improvement factor saturates near $n = 1$ (Figs. 3 and 5), only a simple first-order DPCM system was used. The optimum predictor was recomputed for each separate unvoiced and voiced region and the LPC parameters were calculated at a rate of 50 samples per second. The speech was synthesized using the formulation discussed by Atal and Hanauer.[1] After quantization, the parameters were geometrically interpolated (linear interpolation on a logarithmic scale) to allow pitch synchronous resetting of the synthesizer.

The quantizer used in the DPCM coding of the synthesis parameter was a nonuniform quantizer that was designed to exploit the properties of each parameter's error signal. An experimental investigation has indicated that the difference signal for pitch, power, and g_1 are most suitably modeled by a zero mean gamma density,

$$P_{e_r}(e_r) = \frac{\sqrt{k}}{2\sqrt{\pi |e_r|}} \exp\left(-k|e_r|\right), \tag{9}$$

where

$$\sigma = \frac{\sqrt{0.75}}{k}.$$

The higher-order log-area coefficients are more Laplacian in character:

$$P_{e_r}(e_r) = \frac{1}{2\beta} \exp\left(-\frac{|e_r|}{\beta}\right),$$

where

$$\sigma = \sqrt{2}\beta.$$

A signal with a gamma distribution is highly concentrated near its mean, but can also readily achieve values more than three standard deviations from its mean. A Laplacian signal is less concentrated than a gamma signal near its mean value. Figure 6 illustrates the statistical characteristics of a zero mean, unit standard, deviation signal with a gamma density, a Laplacian density, and a gaussian density. Figure 7 shows a comparison between the calculated distributions for the difference signal of several typical synthesis parameters and their approximated distributions.

For a gamma-behaved signal, the properties of the optimum quantizer are summarized in Table I.[9] The x_i values in the table define the ends of quantizer input ranges, and the y_i values are the corresponding outputs. Thus, for a two-bit quantizer, an input between 0 and 1.205 is quantized as 0.302. Similarly, an input between 0.229 and 0.588 for a four-bit scheme is quantized as 0.386. The properties of the optimum quantizer for Laplacian signals are summarized in Table II.[9] Included in these tables is the expected mean square between the difference

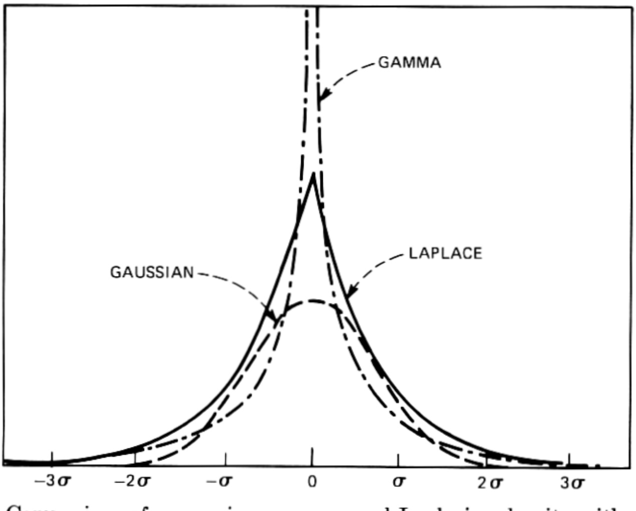


Fig. 6—Comparison of a gaussian, gamma, and Laplacian density with zero mean and unit standard deviation.

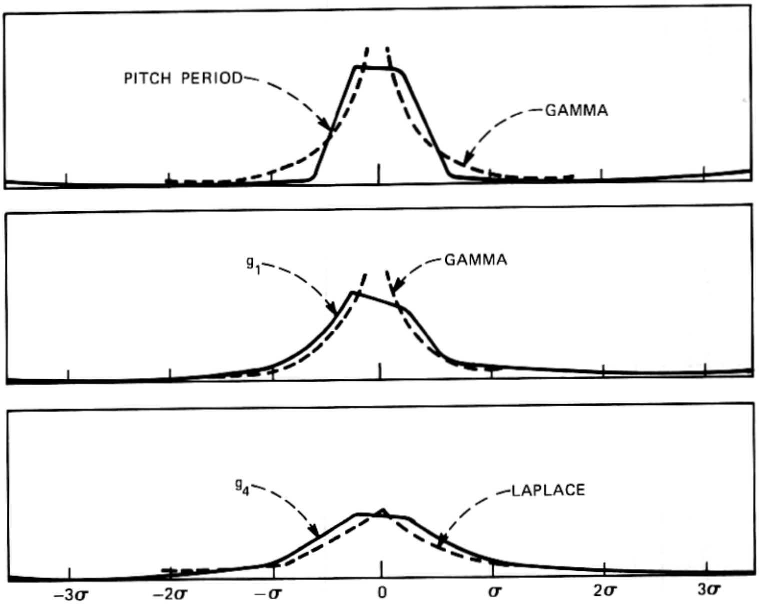


Fig. 7—Comparison between calculated density and approximated density for difference signals.

signal and the quantized difference. Thus, for a four-bit quantization of a gamma signal, the mean square error is 0.0196.

Tables I and II are constructed for signals with unit standard deviation. To obtain the levels y_i and boundaries x_i for signals with standard

Table I — Optimum quantizers for signals with gamma density ($\mu = 0,\ \sigma^2 = 1$)

B	1		2		3		4		5	
i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
1	∞	0.577	1.205	0.302	0.504	0.149	0.229	0.072	0.101	0.033
2			∞	2.108	1.401	0.859	0.588	0.386	0.252	0.169
3					2.872	1.944	1.045	0.791	0.429	0.334
4					∞	3.779	1.623	1.300	0.630	0.523
5							2.372	1.945	0.857	0.737
6							3.407	2.798	1.111	0.976
7							5.050	4.015	1.397	1.245
8							∞	6.085	1.720	1.548
9									2.089	1.892
10									2.517	2.287
11									3.022	2.747
12									3.633	3.296
13									4.404	3.970
14									5.444	4.838
15									7.046	6.050
16									∞	8.043
MSE	0.6680		0.2326		0.0712		0.0196		0.0052	

Table II — Optimum quantizers for signals with Laplace density ($\mu = 0$, $\sigma^2 = 1$)

B	1		2		3		4		5	
i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
1	∞	0.707	1.102	0.395	0.504	0.222	0.266	0.126	0.147	0.072
2			∞	1.810	1.181	0.785	0.566	0.407	0.302	0.222
3					2.285	1.576	0.910	0.726	0.467	0.382
4					∞	2.994	1.317	1.095	0.642	0.551
5							1.821	1.540	0.829	0.732
6							2.499	2.103	1.031	0.926
7							3.605	2.895	1.250	1.136
8							∞	4.316	1.490	1.365
9									1.756	1.616
10									2.055	1.896
11									2.398	2.214
12									2.804	2.583
13									3.305	3.025
14									3.978	3.586
15									5.069	4.371
16									∞	5.768
MSE	0.5		0.1765		0.0548		0.0154		0.00414	

deviation different from unity, simply multiply the given values by the actual standard deviation.* The standard deviation for each parameter can be approximated as the rms power of the unquantized error signal. The rms value of the unquantized error signal is obtained directly from the calculation of the optimum DPCM predictors and is given by

$$\sigma^2 = C_0 - \sum_{i=1}^{n} b_i C_i.^\dagger$$

4.2 Experimental results

Four sentences were synthesized in the experimentation:

A. Few thieves are never sent to the jug.
B. May we all learn a yellow lion roar.
C. It's time we rounded up that herd of Asian cattle.
D. Should we chase those young outlaw cowboys?

High-quality recordings of these sentences were made by two male and two female speakers, and these utterances were used to obtain the analysis data for the DPCM coding method.

* To obtain the mean square error, multiply the values by the signal variance.

† Since the properties of the unquantized error signal are explicitly known, it is sometimes advantageous to use a more complex nonuniform quantizer to truly optimize the transmission system.

Various schemes were tested for assigning bit rates for each individual error signal. From informal listening experiments, it was determined that synthetic speech that was negligibly different from the unquantized synthesis could be generated according to the following bit assignment:

Pitch:	3 bits/frame
Power:	3 bits/frame
Unvoiced-voiced:	1 bit/frame
g_1:	4 bits/frame
g_2:	4 bits/frame
g_3:	4 bits/frame
g_4:	4 bits/frame
g_5:	3 bits/frame
g_6:	3 bits/frame
g_7:	2 bits/frame
g_8:	2 bits/frame
g_9:	2 bits/frame
g_{10}:	1 bit/frame
g_{11}:	1 bit/frame
g_{12}:	1 bit/frame

The total number of bits dedicated to the complete set of LPC parameter is only 38 bits/frame or 1900 bits/s. On the average, an additional 100 bits/s are required to transmit the necessary DPCM information (DPCM predictors, standard deviations, and initial values of the LPC parameters). As can be observed from Figs. 8, 9, and 10, the spectrogram of the DPCM synthetic speech closely resembles that of the unquantized synthetic speech but requires only a fraction of the storage.

As the bit rate for the DPCM linear prediction vocoder is lowered below the value of 2000 bits/s, the quality of the synthesis slowly begins to deviate from that of the unquantized synthesis. Since the log-area parameters are approximately ordered in terms of their sensitivity, the most expandable bits are those allotted to the lower-ordered g_i's.[4] Depending on the speaker and the utterance, the bit rate can be lowered to between 900 and 1500 bits/s and still allow acceptable quality synthesis.* Figures 11, 12, and 13 illustrate the above examples for a bit-rate of 1400 bits/s (3; 3; 1; 4, 3, 2, 2, 1, 1, 1, 1, 1, 1, 1). The synthetic speech in these examples is slightly degraded from the unquantized synthesis, but the speech is still readily understood and the vocal attributes of the speaker are still apparent. It should be appre-

* Acceptable quality speech synthesis is defined as speech containing all the information content of the original without containing any annoying degradation in speech quality.

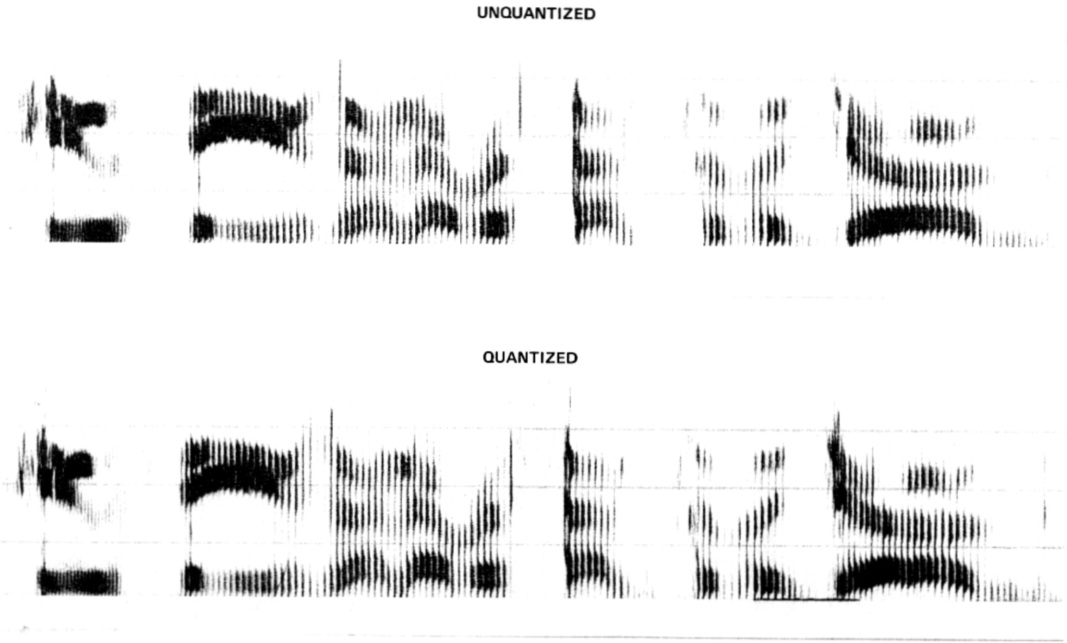


Fig. 8—2000-bits/s quantization of "Few thieves are never sent to the jug." (Male speaker, LG.)

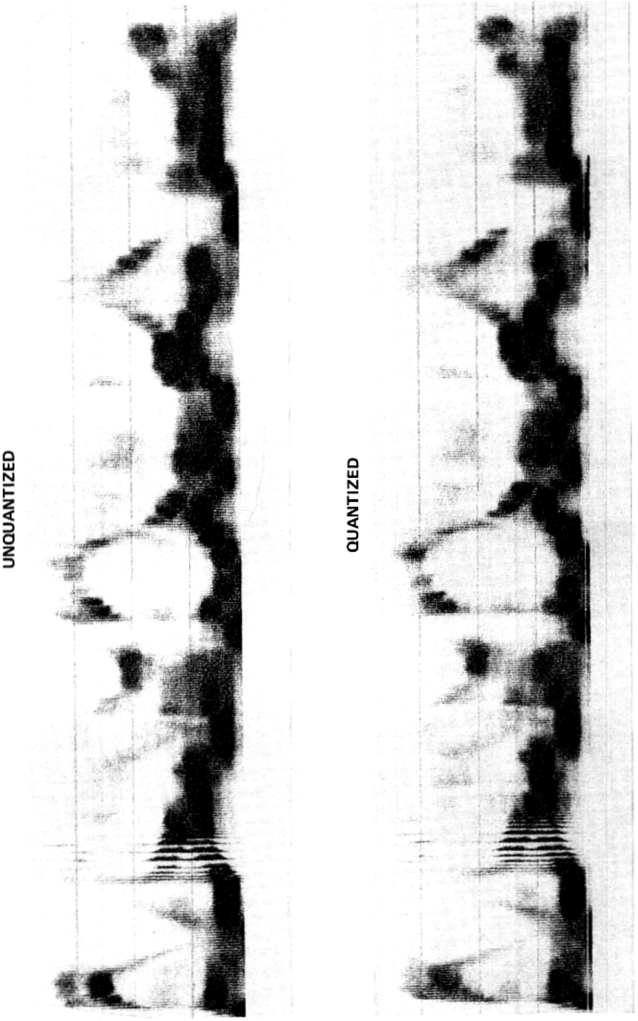


Fig. 9—2000-bits/s quantization of "May we all learn a yellow lion roar." (Female speaker, BG.)

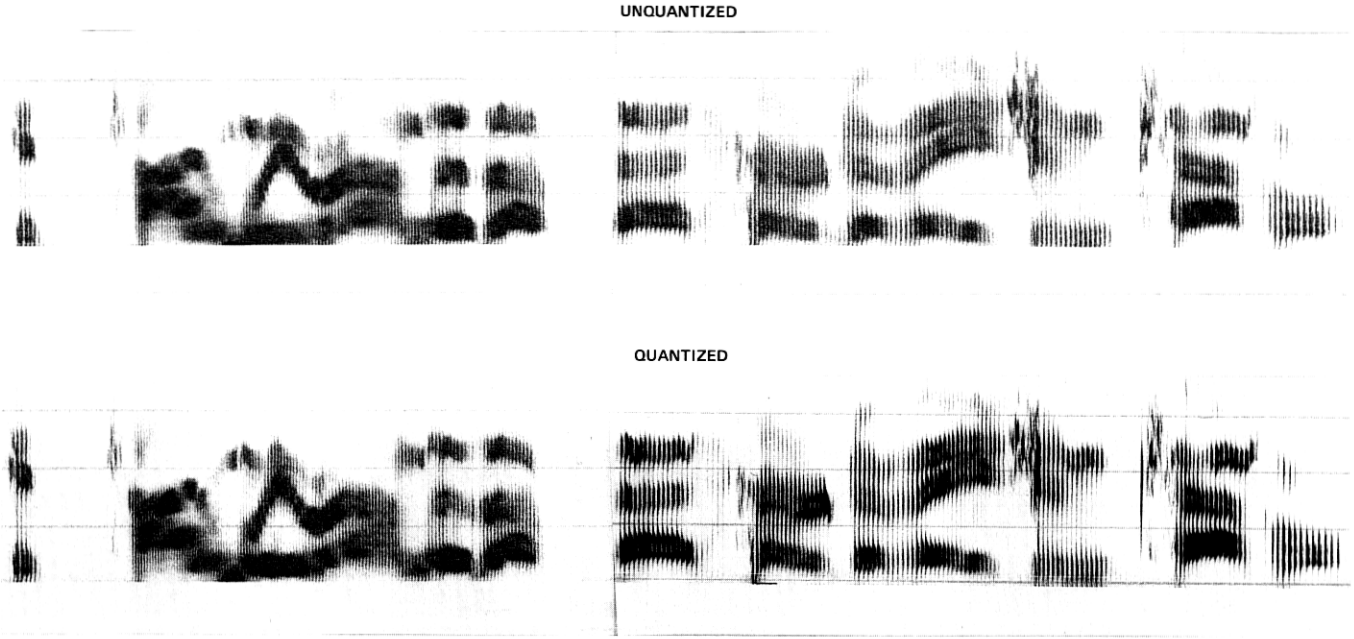


Fig. 10—2000-bits/s quantization of "It's time we rounded up that herd of Asian cattle." (Male speaker, PB.)

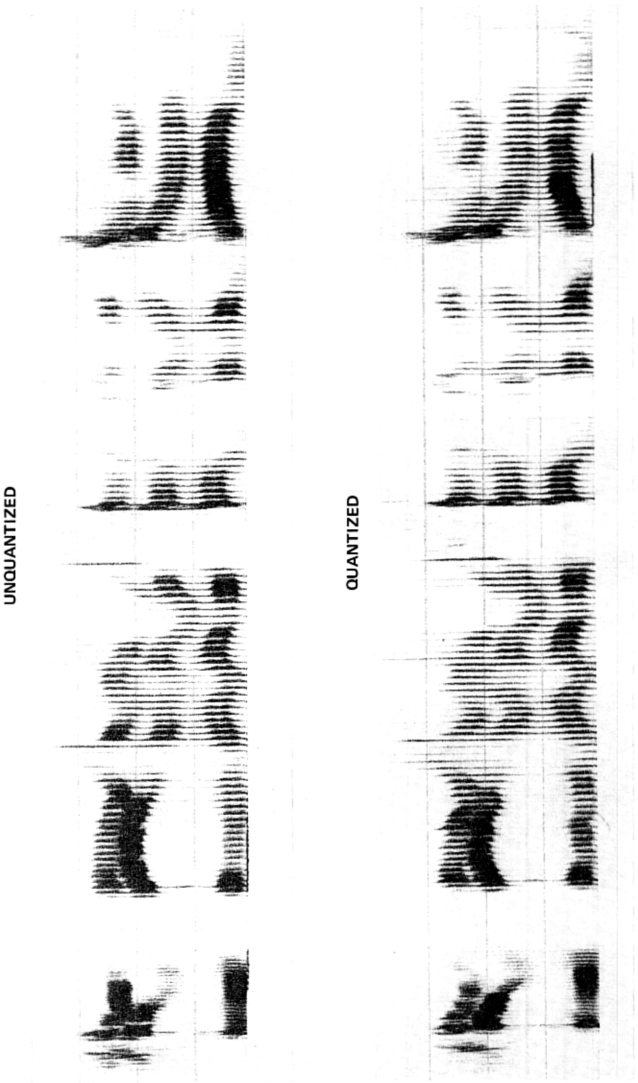


Fig. 11—1400-bits/s quantization of "Few thieves are never sent to the jug." (Male speaker, LG.)

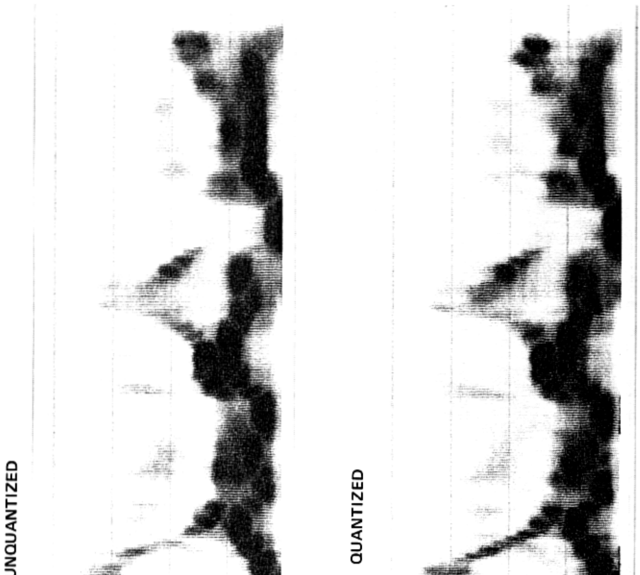


Fig. 12—1400-bits/s quantization of "May we all learn a yellow lion roar." (Female speaker, BM.)

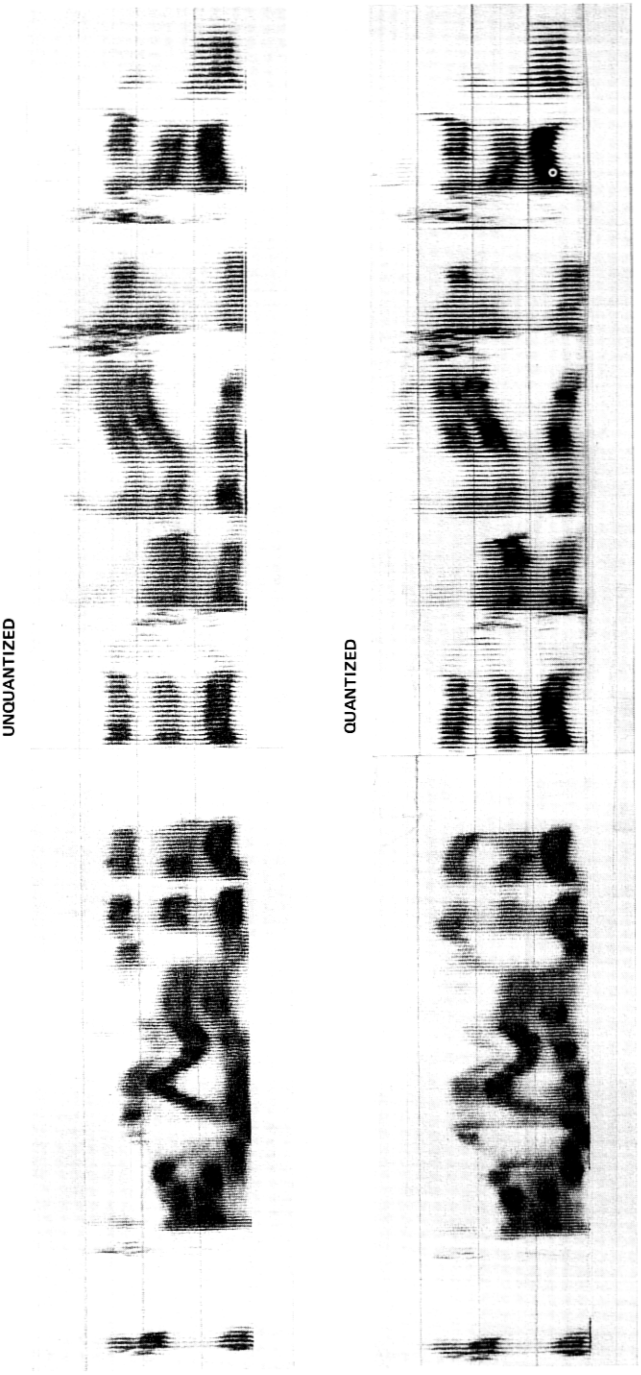


Fig. 13—1400-bits/s quantization of "It's time we rounded up that herd of Asian cattle." (Male speaker, PB.)

ciated that the necessary storage requirements to produce acceptable quality synthetic speech in this method are nearly $\frac{1}{3}$ the requirement for the PCM transmission of the LPC parameters (see Section I).

V. REAL-TIME DPCM TRANSMISSION

The DPCM scheme developed in Section III suffers from the drawback that the calculation of the DPCM predictors and the quantizer step size are delayed until all the LPC parameters are available. For real-time speech synthesis, it is desirable that the process of parameter transmission be done concurrently with the measurement of the LPC parameters. In this section, we discuss several schemes for achieving real-time transmission while still retaining almost the performance of the optimum DPCM strategy.

5.1 *Average statistical system*

The first means of obtaining a real-time system is based upon the observation that the optimum DPCM first-order predictor for many of the LPC parameters is nearly equal to one [$b_1 = 1.0$ in eq. (4)]. Thus, the optimum linear prediction of the parameter x_r is approximately x_{r-1}. Table III is a comparison of the improvement factors G_{opt} obtained for $b_1 = 1.0$ and b_1 set equal to the optimum value. The overall improvement factors for $b_1 = 1.0$ are not significantly different from the optimum values, and the delay in calculating the optimum b_1 can be avoided by simply letting $b_1 = 1.0$.

To design the optimum quantizer, it is necessary to know the standard deviation of the signal to be quantized. However, our statistical studies have indicated that the standard deviation of the various difference signals are quite stable across different utterances and different speakers. Table IV shows the measured standard deviations for each difference signals computed with $b_1 = 1.0$. Table IV also

Table III — Comparison of G_{opt} in decibels with b_1 set equal to optimum value and $b_1 = 1.0$. Sentence A is "Few thieves are never sent to the jug" and sentence B is "May we all learn a yellow lion roar."

	Pitch	Power	g_1	g_2
Sentence A				
b_1 = Optimum	23.7	12.2	14.7	12.2
b_1 = 1.0	20.2	10.1	14.1	11.0
Sentence B				
b_1 = Optimum	33.8	19.0	24.0	19.6
b_1 = 1.0	33.1	18.8	23.9	19.2

Table IV — Measured standard deviations for the synthesis parameters

	Updated	No Updating
Pitch Period	13.01	16.5
Power	27×10^5	27×10^5
g_1	0.697	0.959
g_2	0.729	0.830
g_3	0.509	0.559
g_4	0.510	0.554
g_5	0.413	0.446
g_6	0.417	0.430
g_7	0.386	0.406
g_8	0.385	0.406
g_9	0.377	0.399
g_{10}	0.346	0.364
g_{11}	0.332	0.342
g_{12}	0.322	0.328

contains the standard deviation for a system in which the prediction scheme is not updated for each unvoiced and voiced region.

Using the standard deviations listed in Table IV and the quantizer discussed in Section IV, a robust transmission scheme is achieved. For example, the difference signal for the pitch period can be accurately quantized for differences as small as two samples or as large as 50 for three-bit quantization.* The synthetic speech quality for the average statistical system compares quite favorably to the optimum scheme, and has the added advantage of real-time implementation.

5.2 Adaptive system

5.2.1 Adaptive DPCM prediction

The DPCM predictors can also be calculated without knowing the entire sequence of parameters by an adaptive method that is based upon the technique of "steepest descent."[11] In this scheme, an initial estimate of the DPCM predictors is determined and then a new set of predictors is calculated to reduce the prediction error. The perturbation in the predictors is in a direction opposite the gradient of the prediction error taken with respect to the DPCM predictor vector. The resulting perturbation is given by

$$\delta^r(b_j) = B \cdot \text{sgn}\,(e_r) \cdot \tilde{x}_{r-j} / \sum_{k=1}^{n} |\tilde{x}_{r-k}|, \tag{10}$$

where B is the adaptation rate (typically, $B = 0.09$), and $\tilde{x}_r$ is the

* If a nonlinear smoothing algorithm (Ref. 10) is applied to the raw pitch measurements, the variance of the corresponding difference signal is reduced by more than $\frac{1}{2}$. A two-bit quantization can then be used for pitch without diminishing the quality of the synthesis.

quantizer value of the parameter. For the prediction of the $(r+1)$th sample of the parameter, the DPCM predictors are given by

$$b_j^{r+1} = b_j^r + \delta^r(b_j). \tag{11}$$

For a quantizer with $B \geqq 2$, it can be shown that the adaptation scheme will match the changing spectral properties of the speech signal and result in near-optimum performance.[12] For the two methods given above, it should be noted that, in addition to the on-line calculation of the DPCM predictors, it is unnecessary to transmit the predictors.

5.2.2 Adaptive quantization

In the previous section, the quantizer was constructed to take advantage of the known properties or average statistical properties of each parameter's difference signal. In this part of the paper, we introduce an alternate technique for estimating the signal variance. This method is based upon an adaptive approach developed by Cummiskey, Jayant, and Flanagan.[13] In their scheme, a simple uniform quantization of the difference signal is used, but the step size for every new input is varied by a factor depending on which quantizer slot was occupied by the previous sample. Numbering the quantizer slots in the manner shown in Fig. 14, the updated step size Δ_{r+1} is calculated from the previous step size Δ_r by

$$\Delta_{r+1} = \Delta_r \cdot M(|H_r|), \tag{12}$$

where $H_r = 1, 2, \cdots, B$ and the multiplier function $M(\ )$ is a time-invariant function of the quantizer slot number.

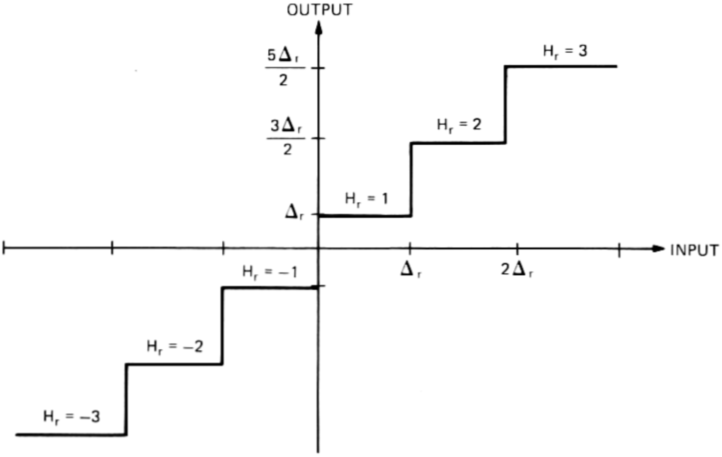


Fig. 14—Numbering of quantizer slots for adaptive quantization.

To adequately match the step size to the signal variance, the multiplier function must be properly chosen. Table V shows the multiplier functions found to be experimentally optimum for quantizing speech waveforms. Using this adaptive scheme (ADPCM) and these multipliers, the quantization of the difference signals can also be efficiently achieved even when the initial step size is a poor estimate of signal variance. Table VI is a comparison of the signal-to-noise ratio for the adaptive scheme with a crude initial estimate of step size and the optimum quantizer discussed in Section IV. The results in Table VI are an encouraging demonstration that it is not necessary to know the statistical structure of the difference signal to efficiently quantize the signal. In fact, it can be shown that, if the properties of the signal are nonstationary, the adaptive method is more suitable than the scheme used in Section IV.

It should be noted that the above scheme does not apply for one-bit quantization ($B = 1$). A simple strategy for one-bit quantization has been developed by Jayant.[14] Let c_r and c_{r-1} denote the values of successive bits in a one-bit scheme, then

$$\Delta_r = \Delta_{r-1}P^{c_r c_{r-1}}, \tag{13}$$

where P has the typical value $P = 1.5$. Although this method was developed for quantizing speech waveforms, it performs quite well in quantizing the parameter difference signals. A comparison of this method and the optimum technique is shown in Table VII. Again, the adaptive scheme works well even with a poor initial estimate of signal variance.

5.3 Synthesis

To subjectively evaluate the performance of the adaptive methods suggested in this section, several speech utterances were synthesized. The synthesis scheme was again the one described by Atal and Hanauer,[1] but an adaptive quantizer and a second-order adaptive predic-

Table V — Step size multipliers for $B = 2$, 3, and 4 (Ref. 7)

	2	3	4
$M(1)$	0.80	0.90	0.90
$M(2)$	1.60	0.90	0.90
$M(3)$		1.25	0.90
$M(4)$		1.75	0.90
$M(5)$			1.20
$M(6)$			1.60
$M(7)$			2.00
$M(8)$			2.40

Table VI — Comparison of the signal-to-noise ratio for the adaptive quantizer with crude initial estimate of step size and the optimum gaussian signal uniform quantizer. The analysis is for the sentence, "May we all learn a yellow lion roar."

Bits	g_1		g_2		g_3	
	Adaptive	Optimum	Adaptive	Optimum	Adaptive	Optimum
2	12.6	13.6	18.3	18.4	15.6	16.5
3	18.0	20.4	21.8	21.8	19.2	20.0
4	22.8	21.9	24.9	23.9	24.0	23.1

tion DPCM technique was used to transmit the LPC parameters. The initial estimates of the predictors were $b_1 = 1.0$ and $b_2 = 0.0$. A second-order analysis was performed because adaptive prediction makes the G_{opt} function saturate at a larger value than a nonadaptive predictor.[7] The initial estimate of the quantizer step size was set equal to the standard deviations of the parameters listed in Table IV. For parameters in which the quantizer uses only one bit, a first-order system with $b_1 = 1.0$ was used.

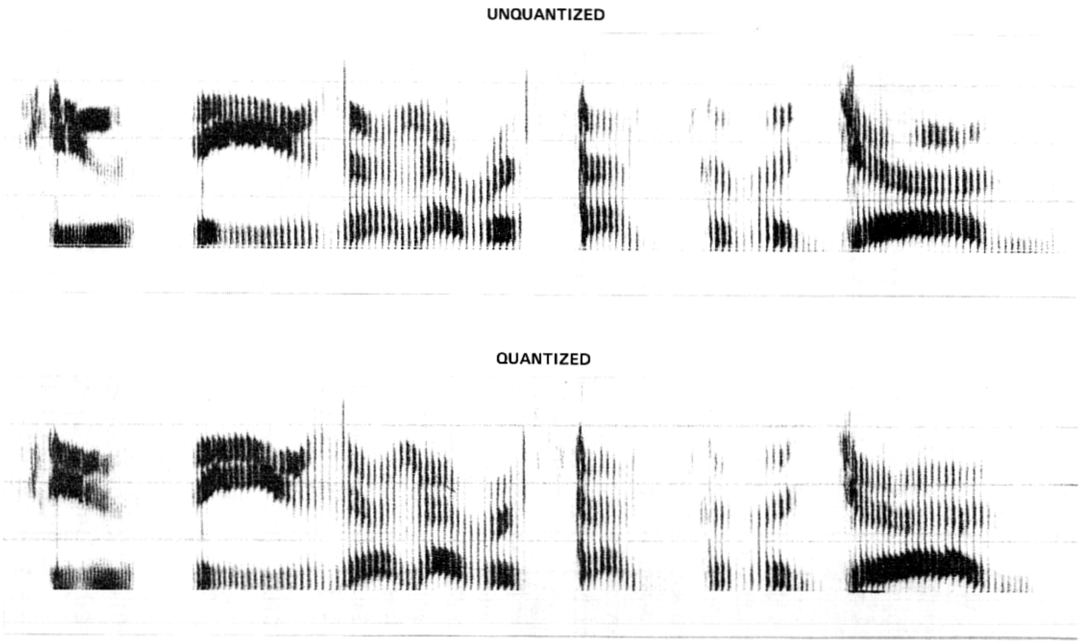
Employing the bit assignment cited in Section IV, the quality of the synthetic speech was only slightly worse than the optimum scheme. Figure 15 shows a comparison of one example of the optimum scheme and the adaptive method. To achieve the performance of the optimum scheme, it has been found necessary to allot approximately one bit more per frame to the most sensitive parameters (usually pitch and power).

VI. ORTHOGONAL LINEAR PREDICTION

In the DPCM method of transmission, the value of the parameter x_r is predicted from previous values of the given parameter. However,

Table VII — Comparison of the signal-to-noise ratio for a one-bit adaptive quantizer and optimum one-bit gaussian signal uniform quantizer

g_1		g_2		g_3	
Adaptive	Optimum	Adaptive	Optimum	Adaptive	Optimum
7.3	8.5	8.8	9.9	8.2	9.7



UNQUANTIZED

QUANTIZED

Fig. 15—Adaptive quantization (2000 bits/s) of "Few thieves are never sent to the jug." (Male speaker, LG.)

the LPC parameters have been experimentally determined to be quite redundant.[15] Thus, the parameter x_r can be predicted not only from its own past values but also in some sense from the values of the *other* LPC parameter. A more efficient method of transmission can then be obtained by exploiting all the available information about a given parameter.

One means of exploiting the redundancy among the LPC parameters is to generate a set of orthogonal parameters that are linear combinations of the original set. The new parameters are uniquely (one to one) related to the LPC parameters and are calculated to be independent of each other and therefore do not contain any mutual information. If the original parameters are redundant, only a small subset of the orthogonal parameter will demonstrate any significant frame-to-frame variation. The process of obtaining the appropriate orthogonal parameters is referred to as an eigenvector analysis.[16] The orthogonal parameters are termed eigenvectors, and each vector's statistical variance is termed the eigenvalue of the eigenvector.

To determine the eigenvectors, we first calculate the covariance matrix of the log-area parameters **R** across the utterance. If we denote g_{ij} as the ith log-area parameter in the jth frame, then the elements of **R** are

$$r_{ik} = \frac{1}{N-1} \sum_{j=1}^{N} (g_{ij} - m_i)(g_{kj} - m_k),$$

where

$$m_i = \frac{1}{N} \sum_{j=1}^{N} g_{ij}$$

and N is the number of frames in the utterance. Given the covariance matrix, the set of eigenvalues λ_i are found by solving the set of simultaneous equations

$$|\mathbf{R} - \lambda \mathbf{I}| = 0,$$

where **I** is the identity matrix and $|\mathbf{A}|$ denotes the determinant of the matrix **A**. The eigenvectors Φ_i are then found as solutions of the equation

$$\lambda_i \Phi_i = \mathbf{R} \Phi_i.$$

To illustrate the behavior of the LPC parameters and the corresponding orthogonal parameters, Table VIII contains a listing of the typical variance (eigenvalues) of each calculated eigenvector parameter across the four utterances examined. The redundancy in the original log-area coefficients is reflected in the fact that more than 90 percent of the total statistical variance is contained in the first five or six eigenvectors.

Table VIII — Measured eigenvalues for the four sentences analyzed:

A. Few thieves are never sent to the jug.
B. May we all learn a yellow lion roar.
C. It's time we rounded up that herd of Asian cattle.
D. Should we chase those young outlaw cowboys?

	A	B	C	D
1	2.62	2.23	1.75	2.75
2	1.44	0.80	1.29	1.58
3	0.67	0.54	0.85	0.6
4	0.44	0.38	0.52	0.36
5	0.25	0.32	0.31	0.28
6	0.21	0.24	0.17	0.22
7	0.10	0.12	0.13	0.16
8	0.09	0.10	0.10	0.15
9	0.08	0.08	0.08	0.09
10	0.06	0.05	0.06	0.06
11	0.03	0.04	0.04	0.06
12	0.02	0.01	0.02	0.03

The higher numbered orthogonal parameters have a relatively small variance and can therefore be considered essentially constant throughout the utterance. Thus, the total information in the 12 log-area parameters can be effectively represented in the space of only the first six eigenvectors.

The redundancy in the LPC parameters is not surprising in view of the fact that the speech signal can be synthesized with only three formant parameters (F_1, F_2, F_3). Thus, the information contained in the 12 log-area coefficients are effectively duplicated in the space of only three formant parameters. The method of orthogonal linear prediction can be viewed as a constraint technique for squeezing the original parameters into a smaller but informationally equivalent set of parameters. The informationally equivalent set is formed by the most significant orthogonal parameters (significance is measured in terms of the standard deviation, or eigenvalue, of the orthogonal parameters).

Experimental studies of a variety of speech utterances have shown that quite acceptable quality synthesis can be generated by transmitting only the six most significant orthogonal parameters, pitch, and power. The synthesis is performed by calculating the LPC parameters from the transmitted orthogonal parameters and *a priori* knowledge of the average values of the least significant orthogonal parameters. For acceptable quality synthesis, only 22 bits/frame are needed.

The allotment of bits was as follows:

Pitch: 3 bits/frame
Power: 3 bits/frame
Unvoiced-voice: 1 bit/frame
First orthogonal parameter: 4 bits/frame
Second orthogonal parameter: 3 bits/frame
Third orthogonal parameter: 3 bits/frame
Fourth orthogonal parameter: 2 bits/frame
Fifth orthogonal parameter: 2 bits/frame
Sixth orthogonal parameter: 1 bit/frame.

The total transmission storage requirement in this technique is 1100 bits/s for the synthesis parameters, 100 bits/s for the DPCM information, and an initial one-time investment of 240 bits* for the necessary eigenvector information. Figures 16 to 18 illustrate the synthetic speech spectrograms generated by this technique for the examples previously examined. Depending on the speaker and the utterance, the bit rate for the synthesis parameters can be reduced to between 600 and 1000 bits/s and still yield acceptable quality speech. The low bit rate required for orthogonal linear prediction is quite attractive, but unfortunately this method involves a complex eigenvector analysis and a delay in transmission to collect the statistical data necessary for the calculation of the eigenvectors.

VII. SUMMARY AND CONCLUSIONS

The goal of this paper was the development of a more efficient method of transmitting the LPC parameters. One proposed method involved the use of DPCM techniques. In DPCM transmission, we take advantage of the predictability of the parameter from its previous values to develop a more effective transmission scheme. Acceptable quality synthetic speech can be generated with DPCM by allotting between 1000 and 1500 bits/s. This rate of information transmission is significantly better than the bit rates necessary for the conventional PCM methods.

To enhance the practical application of the DPCM system, the methods of adaptive quantization and adaptive prediction were discussed. These methods allow the on-line calculation of the DPCM predictors and quantizer step size. To further decrease the storage re-

* Four bits for the average value of each of the six least significant parameters (24 bits) and three bits for each of the 12 coefficients required to compute each orthogonal parameter from the log-area coefficients ($36 \times 6 = 216$ bits).

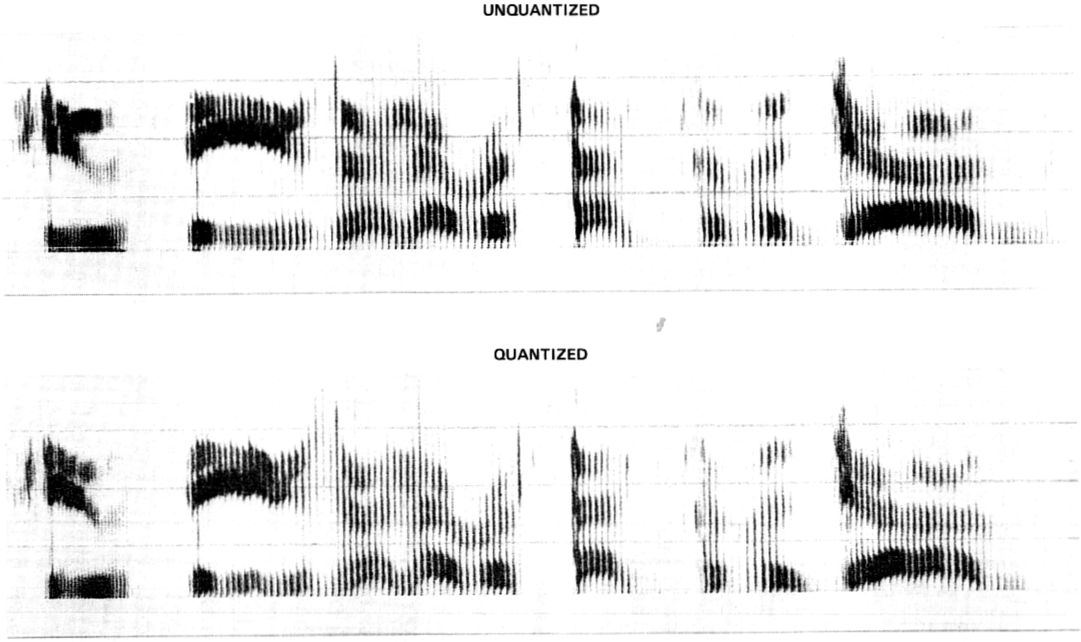


Fig. 16—Quantization of eigenvectors (1200 bits/s) of "Few thieves are never sent to the jug." (Male speaker, LG.)

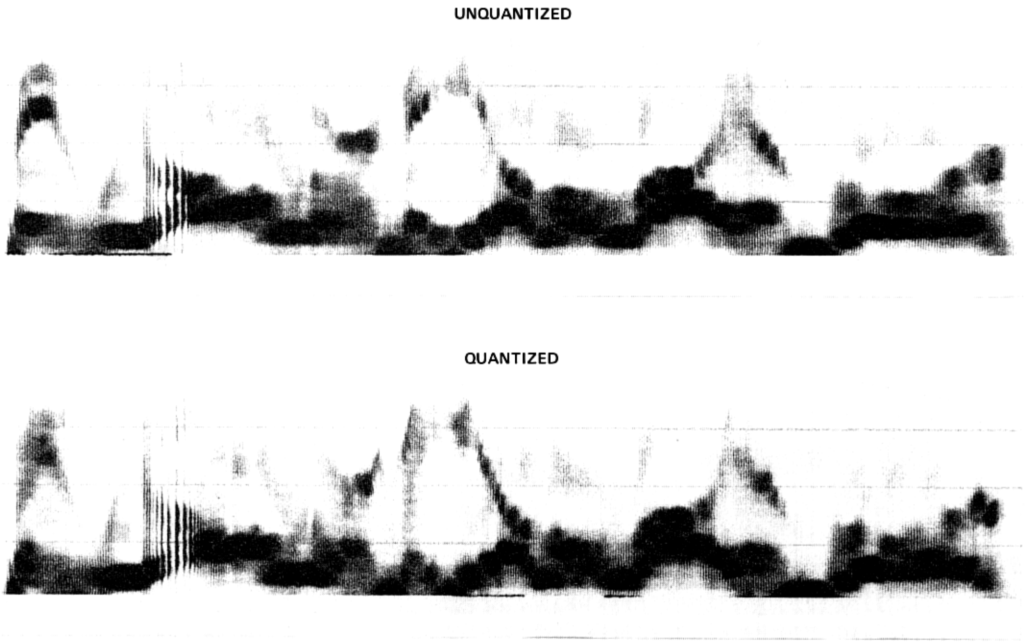


Fig. 17—Quantization of eigenvectors (1200 bits/s) of "May we all learn a yellow lion roar." (Female speaker, BM.)

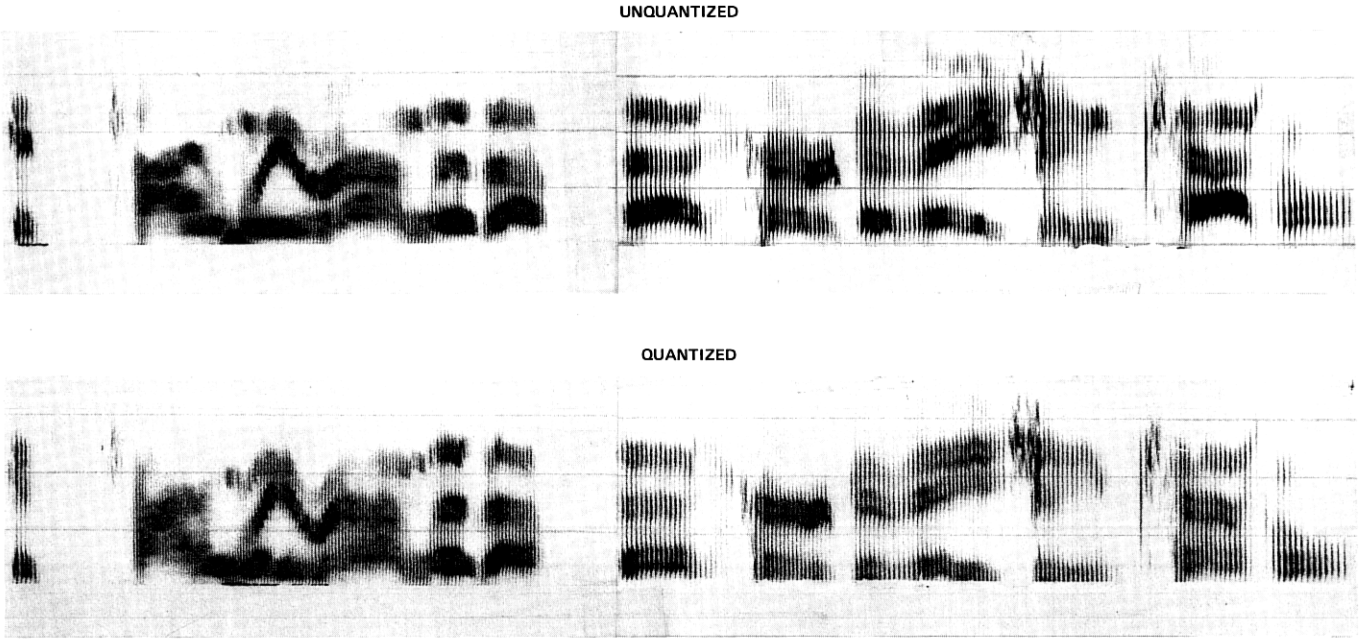


Fig. 18—Quantization of eigenvectors (1200 bits/s) of "It's time we rounded up that herd of Asian cattle." (Male speaker, PB.)

quirements of the LPC vocoder, the redundancy of the log-area parameters was exploited. By transmitting only the most significant eigenvectors, a considerable saving in bit rate can be achieved.

The techniques discussed in this paper are not limited to the transmission of the LPC parameters, but can also be used in conjunction with other vocoder systems. For example, the bit rate of a formant vocoder[4] can be reduced using a DPCM scheme for transmitting the necessary information. These transmission techniques have wide application and can prove very beneficial in a variety of synthesis schemes.

REFERENCES

1. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am., *50*, pp. 637–655, 1971.
2. F. Itakura et al., "An Audio Response Unit Based on Partial Autocorrelation," IEEE Trans. Comm., *COM-20*, No. 4 (August 1972), pp. 792–797.
3. J. D. Markel and A. H. Gray, Jr., "A Linear Prediction Vocoder Simulation Based Upon the Autocorrelation Method," Trans. Acoustics, Speech and Signal Processing, *ASSP-22*, April 1974, pp. 124–134.
4. J. D. Markel, A. H. Gray, Jr., and H. Wakita, "Linear Prediction of Speech Theory and Practice," SCRL Monograph No. 10, Santa Barbara, Cal.: Speech Communications Research Lab, Inc., September 1973.
5. J. L. Flanagan et al., "Synthetic Voices for Computers," IEEE Spectrum, *7*, No. 10 (October 1970), pp. 22–45.
6. R. Viswanathan and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," Cambridge, Mass.: Bolt, Beranek and Newman, Inc., BBN Report No. 2800, April 1974.
7. N. S. Jayant, "Digital Coding of Speech Waveforms: PCM, DPCM, and DM Quantizers." Proc. IEEE, *62*, No. 5 (May 1974), pp. 611–632.
8. B. S. Atal, private communication.
9. M. Paez and T. Glisson, "Minimum Mean Squared Error Quantization in Speech PCM and DCPM Systems," IEEE Trans. Comm., *20* (April 1972), pp. 225–230.
10. L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," to appear in Trans. Acoustics, Speech and Signal Processing.
11. P. Cummiskey, "Adaptive Differential Pulse-Code Modulation for Speech Processing," Ph.D. dissertation, Newark College of Engineering, Newark, N. J., 1973.
12. R. W. Stroh, "Optimum and Adaptive Differential PCM," Ph.D. dissertation, Polytechnic Institute of Brooklyn, Farmingdale, N. Y., 1970.
13. P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., *52*, No. 7 (September 1973), pp. 1105–1118.
14. N. S. Jayant, "Adaptive Delta Modulation with a One-Bit Memory," B.S.T.J., *49*, No. 3 (March 1970), pp. 321–342.
15. A. E. Rosenberg and M. R. Sambur, "An Improved System for Automatic Speaker Verification," presented at the 86th meeting of the Acoustical Society of America, Los Angeles, October 30–November 2, 1973.