AN ADAPTIVE PREDICTIVE DELTA CODER COMBINING SYLLABIC ADAPTATION AND A SELF-ADAPTIVE QUANTIZER

by

Walter Ulrich Schellenberg
Dipl. El. Ing., Swiss Federal Institute
of Technology, Zurich, 1969

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF APPLIED SCIENCE

in the Department
of
Electrical Engineering

We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

April 1974

Department of Electrical Engineering

The University of British Columbia
Vancouver 8, Canada

Date April 30, 1974

Abstract

The purpose of this thesis is to evaluate an adaptive differential encoder for digital communication channels. The redundancy reduction technique used in this coder is not restricted to speech signals only. However, it was optimized for such signals with the objective of keeping the bit-rate as low as possible. In the transmitter, the signal redundancy is reduced in two steps. First, taking advantage of the quasi-periodicity of speech signals, the current signal value is predicted from the value one period before. Secondly, the difference between this prediction and the true value is estimated by a prediction based on the two previous differences. The error of this second prediction is quantized and transmitted to the receiver. The receiver produces a replica of the original signal by adding the received error signal to the predicted value.

The adaptation of the quantizer step size has an exponential characteristic and contains a delay of one sampling period. These two features give rise to instabilities and poor reproduction of high signal frequencies, the latter being an inherent characteristic of this quantizer. The stability problem could not be solved theoretically because the nonlinearity and delay in the quantizer render the system mathematically intractable. By restricting the maximum quantizer level and adding a direct feedback of the step size to the second predictor, the instabilities are restricted to acceptable limits.

The coder was simulated on a digital computer and optimized for sampling rates of 8, 12, and 16 kHz using objective calculations of the signal-to-quantization noise ratio as well as subjective preference tests. In some cases the calculated S/Q ratios allow no distinction in performance, but the subjective evaluations exhibit strong differences in preference. This observation emphasizes the necessity to examine the subjective performance of such voice systems. Comparisons with speech from a log PCM encoder indicate that at a sampling rate of 8 kHz, the subjective quality of the reconstructed speech is slightly superior to that of log PCM encoded at 3 bits per sample and the same sampling frequency. It is estimated that at a sampling rate of 8 kHz, an additional 3800 bits/sec are required to transmit the four coder parameters. The suggested final transmission bit-rate is therefore 11.8 kbits/sec. Thus, a data compression of approximately two to one has been achieved.

Included are suggestions for reducing the bit-rate to 9.6 kbits/sec with minimal degradation in speech quality below that achieved using 11.8 kbits/sec.

TABLE OF CONTENTS

Page

LIST OF ILLUSTRATIONS

Figure Page

ACKNOWLEDGEMENT

I wish to thank Dr. R. W. Donaldson, my supervisor, for suggestions that provided the motivations for this thesis.

I am very thankful to Mr. M. E. Koombes who wrote the programs to digitize the voice signals.

The financial support received in the form of U.B.C. Graduate Fellowships from 1972 to 1974, and under the National Research Council Grant 67-3308 is gratefully acknowledged.

Finally, I would like to thank my wife Ruth for moral support and endless patience while I completed this project.

I. INTRODUCTION

Transmission of speech in digital form offers several advantages over conventional analog techniques. An important benefit of digital channels is their relatively low susceptability to noise and crosstalk. Distorted or weak pulses can be regenerated by repeaters without being cumulatively degraded; therefore maintaining good quality over long distances. Also, some communication links require reliable speech encryption, a rather difficult task in an analog system but digitally realizable with comparatively simple means. Another point that favours digital techniques is the tremendous technological progress in manufacturing digital circuitry, which makes inexpensive, mass produced, and miniaturized subsystems available off the shelf.

The immediate price for these advantages is a higher bandwidth requirement. However, digital techniques also offer ways to compress speech by removal of some redundancy, thus reducing bandwidth. The quality of seven-bit, logarithmic Pulse-Code Modulation (PCM) with a sampling frequency of 8 kHz is generally accepted as good telephone quality. But the necessary bit-rate of 56 kbits/sec compares very unfavourably with the actual information content of speech. The entropy of the written equivalent of speech is only of the order of 50 bits/sec ([1], p. 4). Experiments also indicate that the human is not able to process information at rates in excess of about 50 bits/sec ([1], section 1.3). On the other hand, a conventional analog voice channel requires a bandwidth of at least 3000 Hz and a signal-to-noise ratio of 30 db in order to transmit speech satisfactorily. According to Shannon's theorem such a channel has a capacity of approximately

30,000 bits/sec. Evidently, voice signals contain a lot of redundancy, but their actual information content is not known exactly.

Several systems for transmitting speech intelligibly at low bit-rates have been reported. For a good summary and bibliography see Flanagan [1], chapter VIII, and for some recorded samples see Bayless [2], and Atal [3]. However, high intelligibility is not always sufficient; for example, when the listener would also like to be able to recognize the speaker and the speaker's emotions. Therefore a certain subjective fidelity criterion is specified, and any coding method, although aiming at a reduction of the channel capacity required to transmit the signal, must not impair this criterion severely.

There are mainly two ways to reduce signal redundancy. In the approach taken by vocoders, the speech spectrum is analyzed. The extracted parameters are transmitted and used in the receiver to synthesize a replica of the original signal spectrum. Since they are based on idealized models of speech generation and perception, and discard the information that cannot be parameterized, vocoders usually suffer from unnaturalness of the synthesized speech.

The other approach is predictive coding. As in vocoder systems some characteristics of speech such as pitch and formants are used for a partial parameterization of the signal. But unlike the vocoders, predictive coders transmit the difference between the parameterized signal and the actual input. On the basis of these parameters, the receiver predicts the signal from its past, using the transmitted difference as a correction. Since the entropy of the difference signal is smaller than the entropy of the original signal, it needs less bits for its encoding.

The present thesis deals with predictive coding. The main features of the speech encoders previously published by Atal [6] and Jayant [4] were combined to form a new adaptive predictive coding system.

II. REVIEW OF IMPORTANT PREVIOUS RESULTS

The two systems that are related most closely to the coder described here, and that provided the primary motivations for this project, shall be reviewed briefly. The first section summarizes results obtained by Jayant [4] in a simulation of an adaptive delta modulator. In the second section the speech encoder by Atal and Schroeder [6] using pitch and formant redundancy reduction is described in short.

2.1 An Adaptive Delta Modulator Using a One-Bit Memory

Figure 1 shows this system described by Jayant [4]; its simplicity is very attractive. As in ordinary delta modulation, a two-level quantizer is used. The transmitted channel symbol C_n represents the polarity of the difference between the sampled, bandlimited input signal S_n and the latest approximation to it, P_n. The system is adaptive in the sense that, at every sampling instant nT, the stepsize L_n is modified on the basis of a comparison between the two latest transmitted channel symbols, C_n and C_{n-1}. The adaptation logic is very simple; the new stepsize L_n is calculated by multiplying the latest stepsize L_{n-1} by either +P or -Q, depending on whether C_n and C_{n-1} are equal or not. Conventional delta modulation results for P=Q=1.

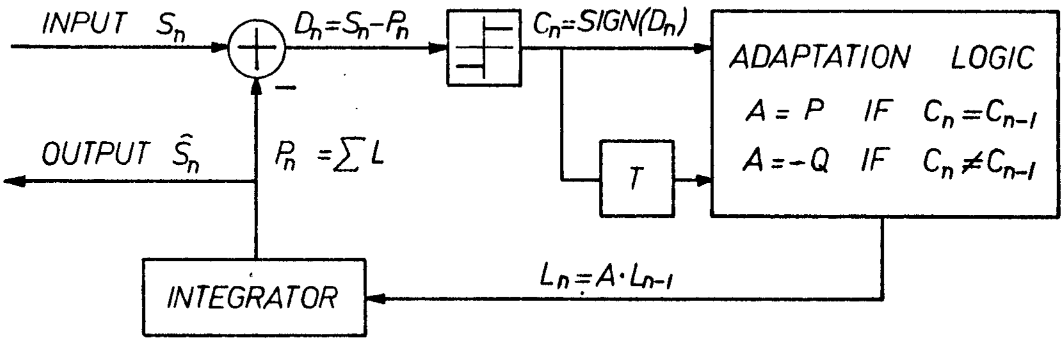


Figure 2.1 Jayant's adaptive delta modulator

2.1.1 Response to a Step Function

Figure 2.2 shows the staircase approximation of a step input. The smallest stepsize is assumed to be 1, P and Q are set at 1.5 and 0.66, respectively. The response has two different phases. In the first phase, the system hunts the input. Due to the exponentially increasing stepsize, this hunting period is considerably shorter than for a conventional delta modulator. In other words, the coder will be less susceptible to slope overload than a nonadaptive scheme. After the staircase approximation has caught up with the input, the oscillating phase starts. The most important feature of this state is the fact that the stepsize does not always assume its smallest possible value. The system may oscillate around the constant input with a nonminimum stepsize, resulting in an increase of the quantization noise during quiet input periods. This is of importance when the minimum allowable stepsize for such a system is chosen.

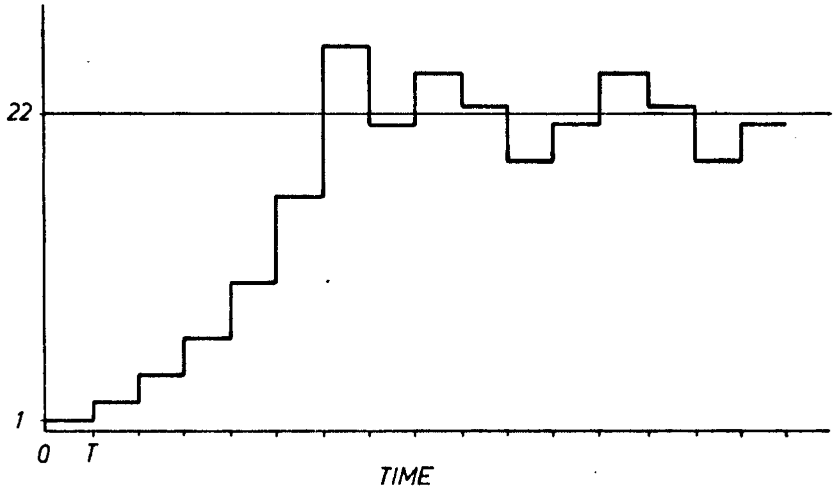


Figure 2.2 Response to a step function.

2.1.2 Bounds on P and Q

Under the assumption that stepsize adaptations using the multipliers P and Q are equally probable, Jayant derived an optimum condition for P and Q, namely:

$$(P\cdot Q)_{opt} = 1. \qquad (2.1)$$

In addition, minimization of the mean square error between input and output yields an upper bound for P_{opt}. Together with the requirement that P<1, his theoretical bounds are

$$1 < P_{opt} = 1/Q_{opt} < 2. \qquad (2.2)$$

When calculating the mean-square error in computer simulations of the coder, using speech as well as video input, the optimum values for P and Q were found to lie within the predicted range; that is,

$$P_{opt} = 1.5$$

$$\text{and} \quad Q_{opt} = 0.6. \qquad (2.3)$$

2.1.3 Performance With Speech Input

Using computer simulations Jayant calculated the signal-to-quantization-noise ratios (S/Q ratios) for the sentence "Have you seen Bill?" for different sampling rates. He found that the adaptive delta modulator outperformed logarithmic PCM at bit-rates lower than 40 kbits/sec. The S/Q ratios obtained were

Sampling rate in kHz	20	40	60
S/Q ratio in db	18	28	34

The subjective optimum differed from these values and occurred for P = 1.2 [5].

2.2 An Adaptive Predictive Coder Using Pitch and Formant Structure

Any efficient source encoder relies on the statistical properties of the signal source. The more the available channel capacity is restricted, the more signal redundancy has to be removed; and therefore the better the signal characteristics have to be known. Unfortunately, the statistics of voice signals are non-stationary. However, the voiced parts of the speech signal contain most of the signal energy and show also the highest correlation between the signal samples [7]. Therefore any redundancy reduction techniques applied to these parts will be most effective. The main sources of redundancy in voiced speech parts are the quasi-periodic nature of the signal and the shape of the spectral envelope. The latter is also called the formant structure and has its origin in acoustic resonances of the vocal tract and the vocal cord. Over short periods of time, such as for the duration of a vowel, the physical shape of the vocal tract hardly changes. Consequently, the formant structure may be considered quasi-stationary. This implies that in order to be optimum, the redundancy reduction process has to be adapted to the momentarily stationary character of speech. Figure 2.3 shows such an adaptive predictor [6].

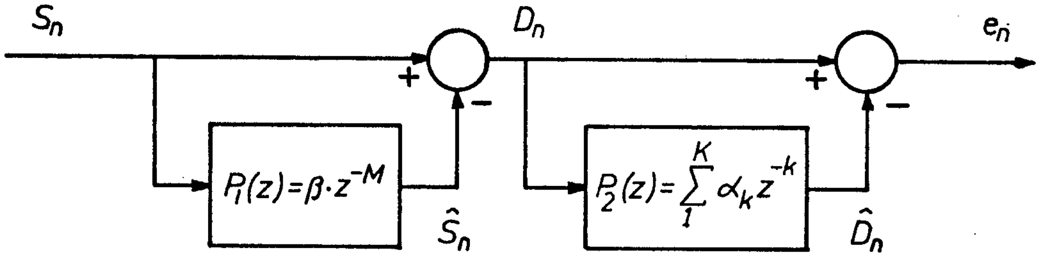


Figure 2.3 Adaptive predictive coder [6].

The input signal S_n is a sample of a bandlimited voice signal. The first loop takes care of the quasi-periodic nature of speech and consists merely of a delay and a gain adjustment. It estimates the present signal value from the value one pitch period before. The second loop forms a linear combination of past values of the first loop output, removing formant information from the spectral envelope [6], [3].

Actually, the predictor error should be minimized by optimizing both loops at the same time. However, this procedure proves mathematically very difficult. Instead, a suboptimum solution can be found by treating the two loops separately. The coefficients for the first predictor $P_1(z)$ are obtained by minimizing the output power of the first loop during a certain period of time N·T for which the predictor is to be optimum. The interval N·T is called the learning period of the system. The output power is related to the sum of the squares of the first loop output, that is,

$$D_n^2 = \sum_{i=1}^{N} (S_i - \beta \cdot S_{i-M})^2 \tag{2.4}$$

The minimum is given by

$$\beta = \frac{\sum\limits_{i=1}^{N} (S_i \cdot S_{i-M})}{\sum\limits_{i=1}^{N} S_{i-M}^2} , \; M = M_{opt} \tag{2.5}$$

where M_{opt} maximizes the normalized correlation

$$\rho = \frac{\sum\limits_{i}^{N} S_i \cdot S_{i-M}}{\left[\sum\limits_{i}^{N} S_i^2 \sum\limits_{i}^{N} S_{i-M}^2\right]^{1/2}} ; \; M > 0. \tag{2.6}$$

This optimization involves considerable computation for each recalculation of the parameter β.

The second loop forms a linear estimate of the difference signal D_n using K previous samples, that is,

$$\hat{D}_n = \sum_{k=1}^{K} \alpha_k \cdot D_{n-k} \qquad (2.7)$$

The mean-square error (MSE) of the predictor output is given by

$$[e_n^2]_{av} = [(D_n - \hat{D}_n)^2]_{av} \; ; \qquad (2.8)$$

Where $[x]_{av}$ stands for the sample mean of x over the learning period N•T. Thus

$$[x]_{av} = \frac{1}{N} \sum_{i=1}^{N} x_i .$$

The MSE can be minimized by setting the partial derivatives of $[e_n^2]_{av}$ with respect to α_k to equal to zero, that is

$$\frac{\delta\, [e_n^2]_{av}}{\delta\, \alpha_j} == [(D_n - \sum_{k=1}^{K} \alpha_k \cdot D_{n-k}) \cdot D_{n-j}]_{av} = 0, \quad j=1,\ldots K. \qquad (2.9)$$

Equation (2.9) yields K simultaneous linear equations for the parameters α_1 to α_K which can be solved using well known computational procedures [8].

Since there are typically three formants and one vocal cord resonance in speech lowpass filtered at 4 kHz, Atal and Schroeder set K equal to 8. In their system, the difference between the predictor output and the true value of the signal was transmitted using a one-bit adaptive quantizer. To minimize the quantization noise, the quantizer level and the predictor parameters β, M, and α_1 to α_8 were readjusted every 5 msec. The computational effort involved in the optimization of the coder is high, but this effort was rewarded by a very good system performance. The authors report subjective comparisons with logarithmic

PCM that rate the quality of their coder only slightly inferior to the quality of a 6-bit log PCM[†]. Both encoders operated at a sampling frequency of 6.67 kHz. No attempt was made to quantize the predictor parameters; but the authors' conjecture is that an additional 3 kbits/sec will suffice to transmit all the parameters.

[†] In another simulation of this scheme by Cummiskey [5], these results could not be duplicated.

III. THE ADAPTIVE PREDICTIVE CODER WITH A SELF-ADJUSTING QUANTIZER LEVEL

This section describes in detail the adaptive delta modulator that was simulated on the IBM 370/168 at the University of British Columbia. The results of the simulations are discussed in chapter V. The scheme implemented combines the two main features of the coders outlined in the previous chapter. Jayant's system is attractive because of its simplicity of implementation. However, the incoming signal has to be highly oversampled in order to be reproduced at the receiver output with an acceptable quality. The other coder yields very good results at a low sampling rate, but requires a large number of computations to remove enough signal redundancy. One objective of this project was to keep the transmission bit-rate of the encoder as low as possible, preferably below 10 kbits/sec. Lowpass filtered speech of telephone quality must be sampled at about 7 kHz. Consequently, in order not to exceed the intended maximum bit rate, only one bit per signal sample may be transmitted (adaptive delta modulation, ADM). The remaining 3 kbits are required to transmit the predictor parameters.

The desired rate of 10 kbits/sec or less makes efficient redundancy reduction a necessity. For this reason, it was decided after some preliminary simulations of a simpler scheme, that the pitch information should also be extracted by the predictor. Consequently, the coder complexity increased considerably.

Some modifications were made to reduce the number of computations required. The second loop was simplified by using only the previous two difference signals D_{n-1} and D_{n-2} instead of the previous eight as in [6] (see also Fig. 2.3). Of course this modification affects the amount of formant information removed by the predictor. However, other studies report only minor improvements when the number of predictor taps is increased from two to eight ([12], p. 130, [3]). Furthermore, to avoid the computations required to optimize the quantizer level for each new interval for which all parameters are recalculated, Jayant's exponential self-adjustment of the level was adopted. This modification also has the advantage that it is unnecessary to transmit parameters to readjust the quantizer level in the receiver. Another modification concerns the first loop; it was found that by inserting a very simple filter the speech quality could be improved noticeably.

Figure 3.1 shows a detailed block diagram of the coder simulated. Note that in order to maintain synchronous tracking, the receiver has to be the exact inverse of the transmitter. The state shown reflects the contents of all registers after the n^{th} sample S_n has entered the coder and the corresponding channel symbol C_n has been transmitted. The system is ready to process the next sample S_{n+1}. The time delays T equal to $1/f_s$, where f_s is the sampling frequency, are included to clarify the sequence of calculations.

3.1 The First Loop: Reduction of the Pitch Redundancy

The gain coefficient β and the delay parameter M were calculated according to (2.5) and (2.6). In order to fit the pitch of different speakers, M was given a range of 20 to 146. This choice was also influenced by the fact that M has to be transmitted to the receiver over a digital channel. It is therefore convenient to let the range of M be a power of 2, minus one.

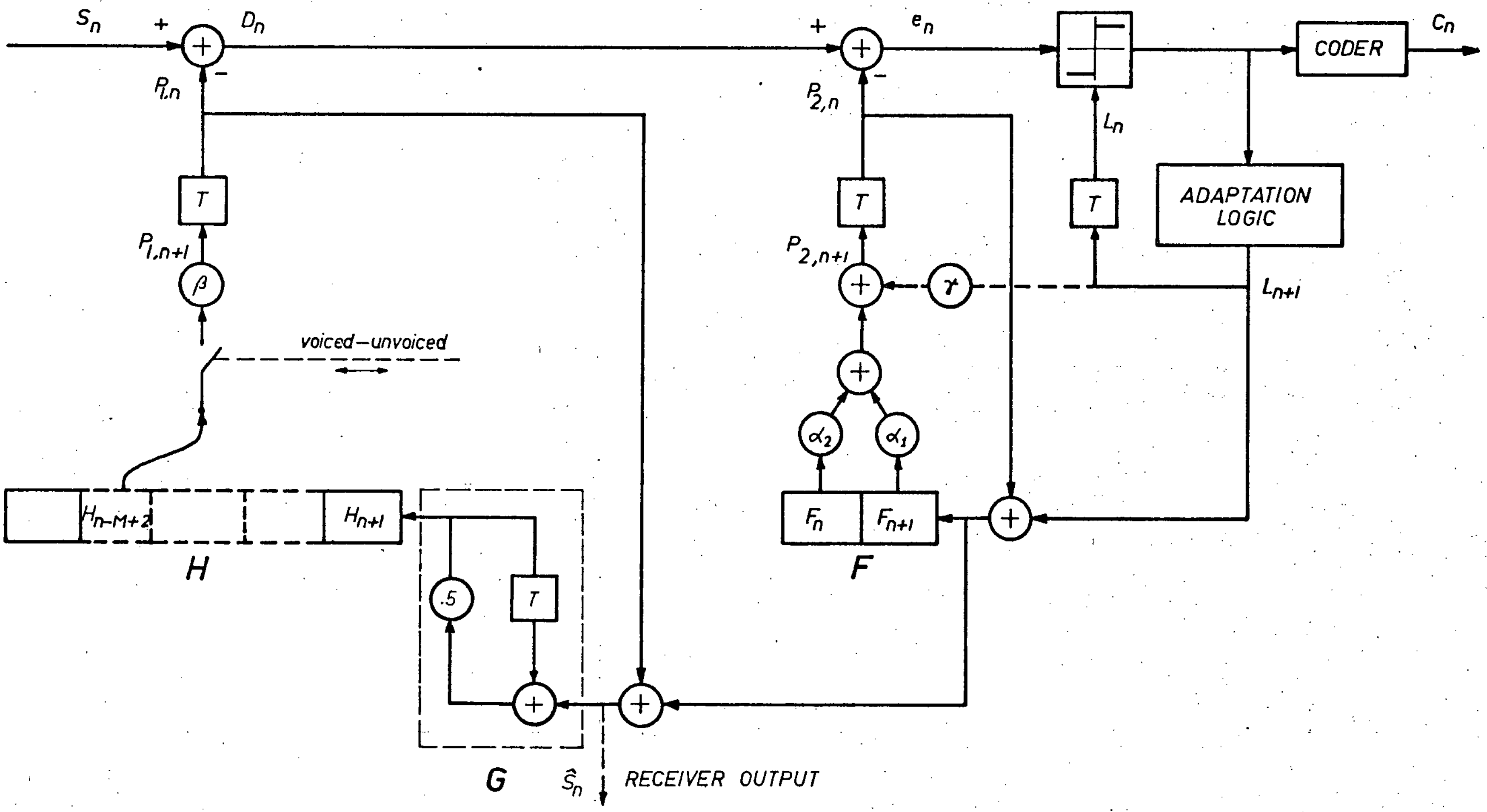


Figure 3.1 Detailed block diagram of the Adaptive Predictive Delta Modulator

Since the purpose of the first predictor loop is to reduce the pitch redundancy, this portion should be operative only during voiced stretches of the input. A switch that sets the new prediction $P_{1,n+1}$ in Fig. 3.1 to zero during unvoiced or quiet periods was added to the loop. The decision voiced/unvoiced is based on the zero-crossing rate of the input signal for the previous learning period. The noise-like, unvoiced fricatives have a much higher zero-crossing rate than the voiced sounds.

During pauses in the input speech β should be set to zero to minimize idle noise generation by the coder itself. This is done by observing the input energy level during the previous learning period. If said energy drops below a certain level, β is set to zero. Even if the input signal energy is small but finite, for example if background noise is present during speech pauses, setting β to zero is acceptable since such signals are not periodic. However, if the receiver were switched off completely, background noise would be absent during short speech pauses, giving the listener the impression of having been cut off from the voice circuit.

The simulations showed that both the zero-crossing rate and the energy level were fairly noncritical and could be varied over a wide range without affecting the performance (see section 5.3.5). The coder complexity is not severely increased by the above modifications. A hardware implementation of the switch is quite simple. The zero-crossings can be detected and counted using a limiter followed by a resettable counter. A rectifier and a resettable integrator are all that are required to determine the energy level.

The second addition to the first predictor loop is the digital filter G preceding the register H. The filter smooths the estimate $\hat{S}_n$ of the input before it is stored in the pitch register H. Signal $\hat{S}_n$ is identical with the receiver output and therefore contains the transmitter input signal distorted by some quantization noise added by the coder. Figure 3.2 shows the relative spectral levels of the input signal ("Joe took father's shoe bench out") and the quantization noise at the receiver output. The spectrum of the noise is, as expected, flatter than that of the input voice signal. Because only a small portion of the entire noise energy lies outside the input signal bandwidth (0 - 3740 Hz), filtering does not improve the signal-to-quantization noise ratio considerably. Nevertheless, other considerations that will be explained in the next paragraph made it advisable to insert a filter into the first loop.

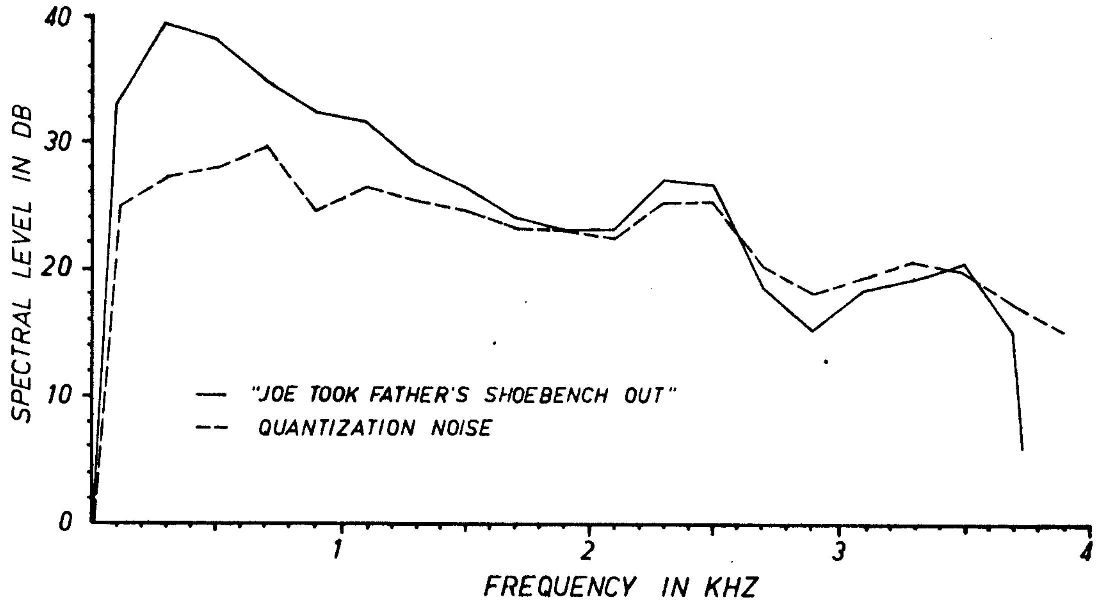


Figure 3.2 Relative spectral levels of a test sentence and the associated quantization noise

The coder approximates the input signal by a staircase with a variable stepsize. If the input contains some fairly rapid changes the approximation is usually much more coarse than for a slowly varying input. The output D_n of the first loop then oscillates even more rapidly, and compared to more quiet periods, at a relatively high amplitude. Since the stepsize is readjusted only for the next prediction, the feedback of the second loop contains a delay, and the predictor cannot keep up with the fast input variations. In these cases it was observed that the stepsize was decreased instead of increased, making the tracking even worse. This situation can be improved only by reducing the error amplitude superimposed on the input signal of the pitch register H; this is the purpose of the filter G.

In realizing filter G, a recursive Chebyshev lowpass filter of fourth order and a cutoff frequency of 3700 Hz was first tried, followed by Butterworth filter of the same type. The results were not very encouraging, in that the S/Q ratio at the receiver output improved by much less than one db. The reason for this is that these filters have an unfavourable impulse response.

Their transfer functions are of the form

$$G(s) = \frac{11}{\prod_{i}[(s+a_i)^2 + b_i^2]} \tag{3.1}$$

with impulse response

$$g(t) = \sum_{i} c_i \cdot e^{-a_i t} \sin(b_i t). \tag{3.2}$$

The response is exponentially decaying, but has some overshoot due to the terms $\sin(b_i t)$. Although the filtered signal will contain less noise, the nature of this noise will be altered. The negative overshoot of the response to a positive input pulse might, for example, add to the negative response to a negative input pulse. Therefore a filter of this type does not really improve the output of the first loop in the sense that it becomes easier to track by the second loop. A filter with an exponentially decaying impulse response and no overshoot seems to be the answer. Such a filter is shown in Figure 3.3. The multiplicative factor a is arbitrary so long as it is less than unity. Experimentally it was found that best results were obtained for a= 0.5.

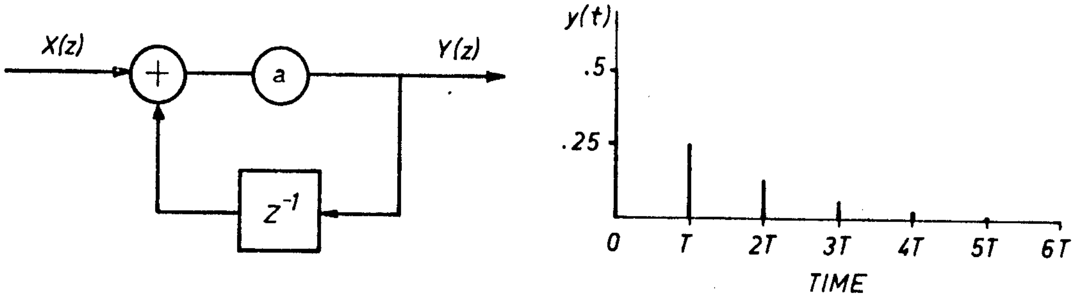


Figure 3.3 Configuration and impulse response to a pulse of height 1 at t=0 of a filter with an exponentially decreasing impulse response and no overshoot

The transfer function of the filter in Fig. 3.3 is readily found using z-transforms.

$$Y(z) = \frac{1}{2}\,[Y(z)\cdot z^{-1} + X(z)]. \qquad (3.3)$$

Hence,

$$Y(z) = \frac{1}{2 - z^{-1}}\,X(z) = H(z)\cdot X(z). \qquad (3.4)$$

To find the frequency response, let $z=e^{j\omega T}$; in which case

$$H(\omega) = \frac{1}{2 - e^{-j\omega T}} \qquad (3.5)$$

The response for frequencies between 0 and 4 kHz is shown in Fig. 3.4. The attenuation curve is almost sinusoidal and has a maximum of 9.5 db.

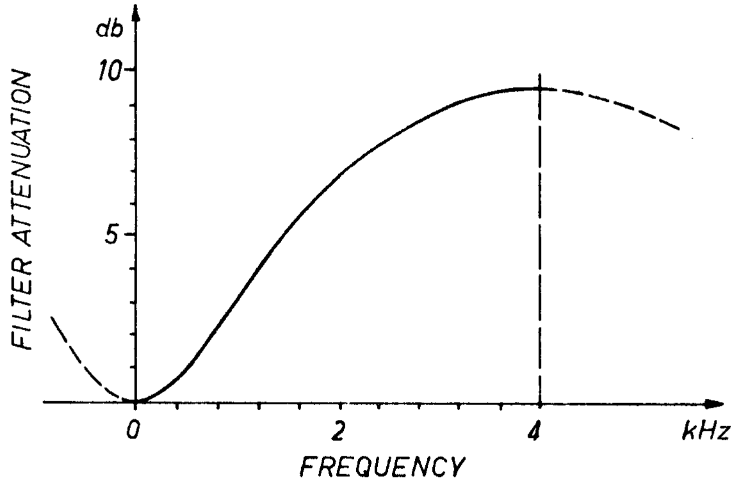


Figure 3.4 Frequency response of the filter in Figure 3.3

The filter is explained best as a device that averages the input over a short period of time. Rewriting (3.3) in the time domain, we find

$$y(nT) = \frac{1}{2}\,[y(nT-T) + x(nT)]. \qquad (3.6)$$

The new filter output is the mean of the old output and the new input. Therefore, no overshoot is possible. Figure 3.5 illustrates this property.

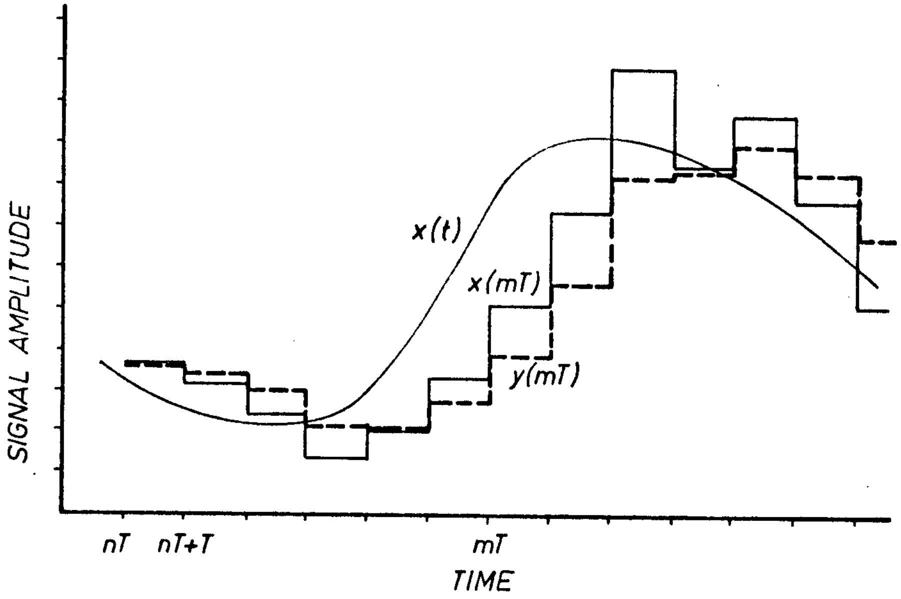


Figure 3.5 Input x(nT) and output y(nT) of the filter in Fig. 3.3

Because of the time lag introduced by any physically realizable lowpass filter, the filtered output is delayed even more with respect to the true signal than is the input. A time shift by T to the left would result in a much better fit. This was also experimentally verified.

Inserting this filter into the first loop and replacing the delay parameter M by M-1 to take into account the above mentioned forward time shift, resulted in a S/Q ratio improvement of 1 to 2 db. The subjective quality also improved noticeably (see section 5.3.1).

Another filter using one previous sample only is given by

$$y(nT) = \frac{1}{2}\,[x(nT-T) + x(nT)] \qquad (3.7)$$

This filter and its impulse response are shown in Figure 3.6.

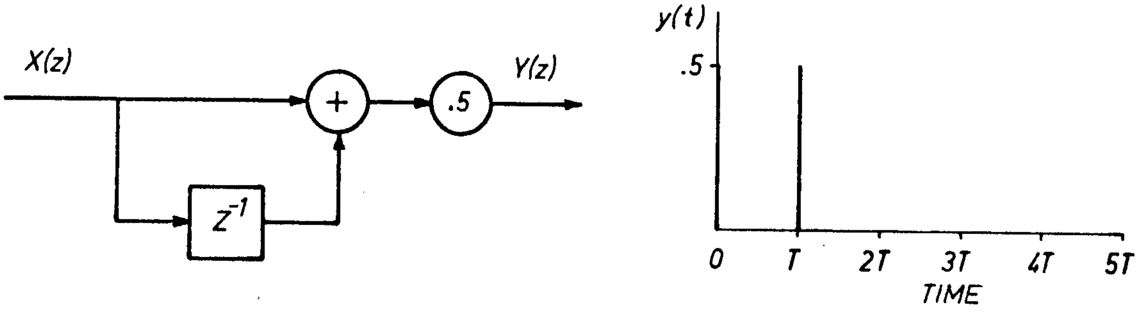


Figure 3.6 Filter with a rectangular impulse response

This filter and several others whose outputs depended on more than just one previous sample were tried, but none of them gave better results than the one described by (3.6). Generally, filters using more than one previous sample have a smaller but more slowly decaying impulse response, i.e. they average the input signal over a longer period of time. Such heavy lowpass filtering reduces the pitch redundancy reduction achieved by the first predictor loop, and is therefore not desirable.

3.2 The Second Loop: Reduction of the Formant Redundancy

The second loop produces a prediction $P_{2,n}$ of the output of the first loop, D_n. The error e_n is then quantized to two levels, encoded, and transmitted as channel symbol C_n. The quantizer level logic is the same as described in Figure 2.1. The two coefficients α_1 and α_2 are calculated using (2.9) with K equal to two. From (2.9) we find

$$\frac{\delta[e_n^2]_{av}}{\delta\alpha_1} = [D_n - (\alpha_1 \cdot D_{n-1} + \alpha_2 \cdot D_{n-2}) \cdot D_{n-1}]_{av} = 0, \qquad (3.8a)$$

and

$$\frac{\delta[e_n^2]_{av}}{\delta\alpha_2} = [D_n - (\alpha_2 \cdot D_{n-1} + \alpha\alpha_2 \cdot D_{n-2}) \cdot D_{n-2}]_{av} = 0, \qquad (3.8b)$$

where $[x]_{av}$ denotes the sample mean of x over the learning period N•T. Equations (3.8a) and (3.8b) can now be solved for α_1 and α_2.

If K previous samples are used, K linear equations for α_1 to α_K are obtained from (2.9). They can be rewritten in matrix notation as

$$\Delta \circ A \cdot\cdot\; \Theta \qquad \Delta \cdot A = \Theta \qquad (3.9)$$

where Δ is a K by K covariance matrix with its (ij)th element given by

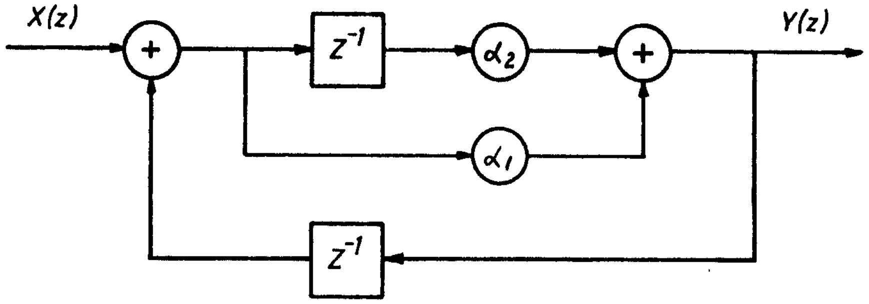
$$\delta_{ij} = [D_{n-i} \cdot D_{n-j}]_{av} = \frac{1}{N} \sum_{n=1}^{N} D_{n-i} \cdot D_{n-j}, \qquad (3.10)$$

A is a column vector of dimension K, containing the coefficients α_j, and Θ is a correlation vector of dimension K whose j^{th} component θ_j is given by

$$\theta_j = [D_n \; D_{n-j}]_{av} = \frac{1}{N} \sum_{n=1}^{N} D_n \cdot D_{n-j}. \qquad (3.11)$$

If (3.9) is solved for A using matrix inversion, a solution is impossible if the matrix Δ is singular. In this case, a unique solution can always be forced by adding a small quantity to each diagonal element of A. For larger values of K, one should take advantage of the symmetry of Δ and apply methods that require less computation and that do not involve matrix inversion [8]. Also it should be noted that ultimately the coefficients α_j will have to be quantized, therefore stimulating the use of iterative techniques to solve (3.9).

The main problem of the second loop is its stability. It is desirable that a feedback system be fully controllable, that is, that each subsystem be stable. In particular, stability is a necessity for the recursive filter in the second loop. Figure 3.7 shows this part of the coder again, where X(z) and Y(z) correspond to L_{n+1} and $P_{2,n+1}$, respectively (see Figure 3.1). The dashed line in Figure 3.1 is omitted and will be discussed later.



Figute 3.7 The recursive filter in the second loop

The requirements for stability can be derived easily using z-transforms. The system equation is

$$Y(z) = \alpha_1[X(z) + Y(z)z^{-1}] + \alpha_2 z^{-1}[X(z) + Y(z)z^{-1}].$$

Hence,

$$Y(z) = \frac{\alpha_1 + \alpha_2 z^{-1}}{1 - \alpha_1 z^{-1} - \alpha_2 z^{-2}} \cdot X(z),$$

or

$$Y(z) = \frac{\alpha_1 z^2 + \alpha_2 z}{z^2 - \alpha_1 z - \alpha_2} \cdot X(z) = H(z)\cdot X(z) \qquad (3.12)$$

In order for the filter to be stable, its transfer function H(z) has to be finite for all $z = e^{j\omega T}$, that is, the poles z_i of H(z) have to lie within the unit circle. Thus, we require

$$|z_i| \leqslant 1 - \varepsilon, \ \varepsilon > 0. \qquad (3.13)$$

The poles z_i are

$$z_{1,2} = \alpha_1/2 \pm[[\alpha_1^2/4 + \alpha_2]^{1/2} = r \cdot e^{\pm ja}, \qquad (3.14)$$

where

$$a = \arg(z_1)$$

and

$$r = |z_{1,2}| = [\alpha_1^2/4 + |\alpha_1^2/4 + \alpha_2|]^{1/2}. \qquad 3.15)$$

If the poles z_i are complex

$$\alpha_1^2/4 + \alpha_2 < 0,$$

and

$$|\alpha_1^2/4 + \alpha_2| = -(\alpha_1^2/4 + \alpha_2). \qquad (3.16)$$

Using (3.16) in (3.15) and observing (3.12), we obtain the following bound for α_2

$$\alpha_2 \geqslant -(1 - \varepsilon). \qquad (3.17)$$

Another restriction is found by selecting real poles $z_{1,2}$ such that $0 \leqslant z_{1,2} \leqslant 1 - \varepsilon$. From (3.14) and (3.13), we obtain

$$\alpha_1/2 \pm [\alpha_1^2/4 + \alpha_2]^{1/2} \leqslant 1 - \varepsilon. \qquad (3.18)$$

Re-arranging (3.18), we get two equations

$$[\alpha_1^2/4 + \alpha_2]^{1/2} \leqslant 1 - \varepsilon - \alpha_1/2, \tag{3.19a}$$

and

$$-[\alpha_1^2/4 + \alpha_2]^{1/2} \leqslant 1 - \varepsilon - \alpha_1/2. \tag{3.19b}$$

Since both sides in (3.19a) are positive, (3.19a) may be squared. After reordering, the following bound on α_1 and α_2 is found

$$(1 - \varepsilon)\cdot\alpha_1 + \alpha_2 \leqslant (1 - \varepsilon)^2 \simeq 1 - 2\varepsilon. \tag{3.20a}$$

The third constraint is obtained by assuming real poles $z_{1,2}$ such that $-(1 - \varepsilon) < z_{1,2} < 0$ and following the same steps as above. The result is

$$- (1 - \varepsilon)\cdot\alpha_1 + \alpha_2 \leqslant (1 - \varepsilon)^2 \simeq 1 - 2\varepsilon. \tag{3.20b}$$

A geometrical interpretation of (3.17), (3.20a), and (3.20b) is shown in Figure 3.8a. The dashed lines represent the boundaries of the stability range for α_1 and α_2 as given by (3.17), (3.19a), and (3.19b). Points P initially outside the triangle are moved in a straight line towards the origin, terminating on the stability boundary. The solid lines of the outer triangle are obtained for $\varepsilon = 0$.

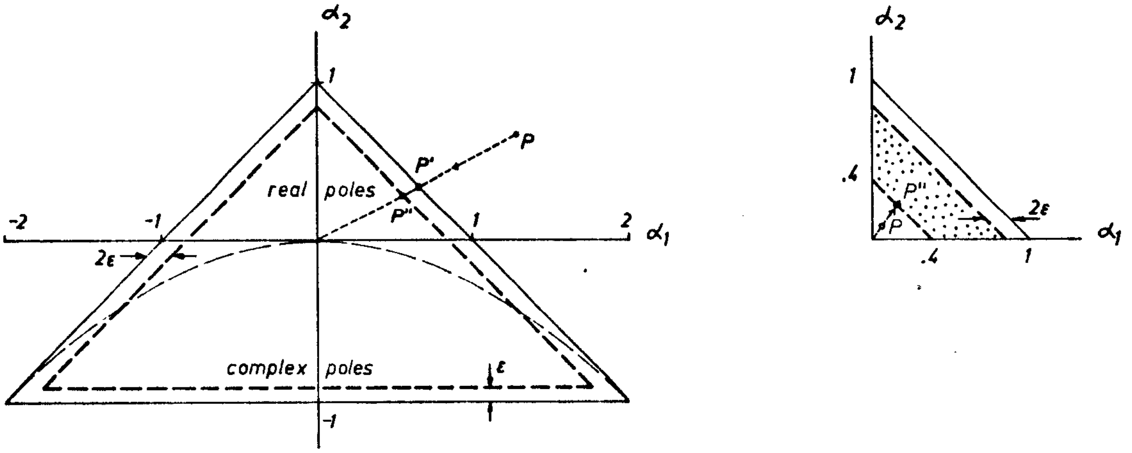


Figure 3.8 a) Stability range for the coefficients α_1 and α_2
b) Truncated stability range

The value of ε can be found as follows. The magnitude of the filter transfer function H(z) is

$$|H(z)| = \frac{|\alpha_1 z^2 + \alpha_2 z|}{|(z - z_o)|\cdot|(z - z_o^*)|} ,$$

where z_o and z_o^* are complex conjugate poles as given by (3.14). The denominator can be interpreted geometrically (see Figure 3.9) as the product of the lengths of the vectors $(z - z_o)$ and $(z - z_o^*)$.

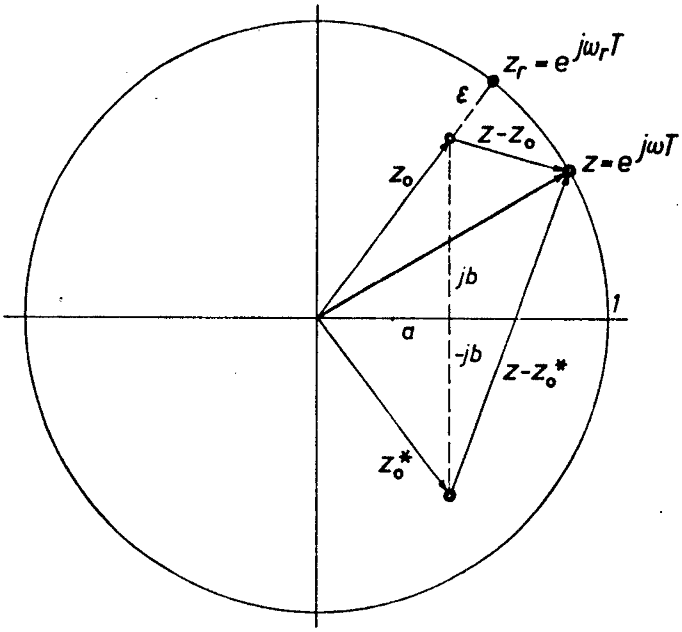


Figure 3.9 Geometrical interpretation of the denominator of (3.8)

The minimum of this product corresponds to the maximum of $|H(z)|$ and occurs for $z = e^{j\omega_r T}$. For this value of z, the length of the vector $(z - z_o)$ is the shortest possible, and equal to ε.

The coefficients α_1 and α_2 are calculated such that the filter transfer function H(z) approximates the spectral envelope of the input signal. In order to avoid some additional degradation of the reconstructed speech, the minimum bandwidth of the filter transfer function should be restricted to a value comparable to that of the bandwidth of spectral peaks of the input signal. Analysis of the spectral envelope of voice signals show that the bandwidth of the first formant usually lies between approximately 40 and 70 Hz ([1], p. 182). In our simulation

$$\varepsilon = 0.024 \tag{3.21}$$

was selected, corresponding to a minimum bandwidth of H(z) of 60 Hz. Other bandwidth restrictions of H(z) can be imposed easily by a suitable choice of ε.

Since the above stability analysis neglects the nonlinearity contained in the second loop, (3.17), (3.20a,b), and (3.21) are not sufficient conditions to guarantee stability for this loop. In fact, it was observed that some instabilities were still present. Figure 3.10 shows this loop again. The quantizer level calculator is represented by the nonlinear part NL.

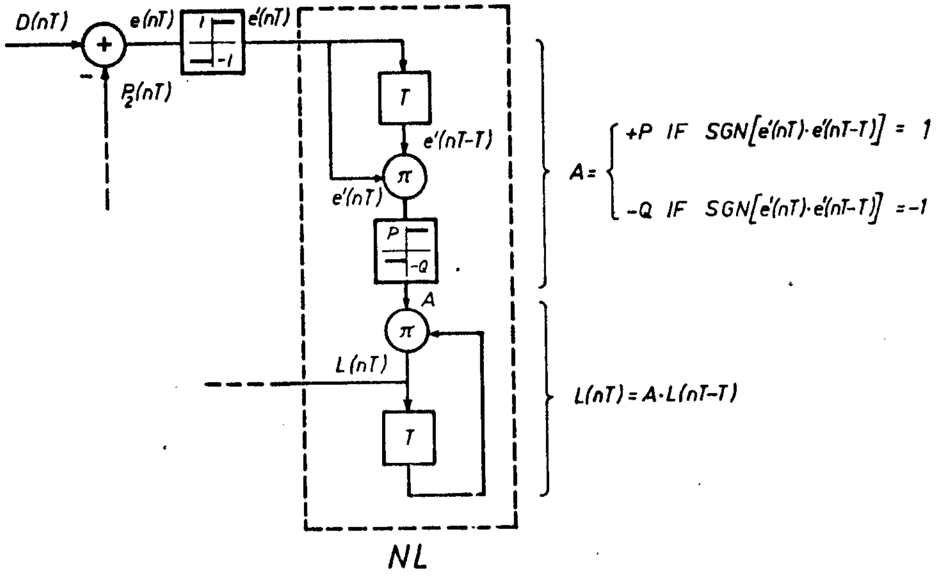


Figure 3.10 The nonlinearity in the second loop

This kind of a nonlinearity with memory makes an analytical solution very difficult. Experimentally, it was found that the system was stable for all coefficients α restricted to the truncated stability region shown in Figure 3.8b by dashed lines. Negative values of α were set to zero, while points outside the area in the first quadrant again moved on a straight line through the origin as shown. The unfortunate result was an almost complete loss of fricatives in the reconstructed speech.

Some negative coefficients, needed to reproduce these sounds, gave rise to instabilities; attempts were made to exclude such coefficient values but no completely satisfactory solution could be found. By limiting the quantizer level to a maximum of 1/8 of the peak signal amplitude the overall error could be reduced considerably; but the larger errors during the instabilities appear as quite audible clicks.

Common to all instabilities was the phenomenon that the coefficients α_1 and α_2 had values such that the prediction P_2 could never reach the input signal D even though the quantizer level was steadily increasing in magnitude. To avoid this situation, a more direct feedback of the quantizer level was introduced into the second loop, as shown by the dashed line in Figure 3.1. Now, since the newly calculated stepsize always has the sign of the difference between the input signal and its prediction, and is increasing exponentially, this additional feedback must eventually catch any run-away situation. How long it takes to reach that condition depends upon the values α_1 and α_2 as well as on the feedback coefficient γ. During this time, the error might still build up considerably, but at least it cannot increase beyond limits. Although adding this direct feedback to the simulation program did not result in a lower S/Q ratio, the subjective impression of the reproduced speech was considerably better (see section 5.3.2). It was also observed that the reproduction of fricative sounds became worse again with increasing values of the feedback coefficient γ.

IV. ON MEASURING AND COMPARING SPEECH QUALITY

In the introduction it was mentioned that the information content of speech is not known exactly, and that high intelligibility of a voice system does not necessarily mean that the subjective quality rating will be equally high. Sound perception is very complex and not yet fully understood. It is not possible to establish a mathematically defined, absolute quality measure that is equivalent to subjective quality evaluation.

To evaluate the performance of a system under development using subjective listening tests exclusively is too tedious. It seems intuitively satisfying that the closer the waveforms of the generated replica are to the true signal, the better the subjective quality will be. A convenient measure for this difference is the mean square error (MSE). The relation between MSE and subjectively rated quality has been studied elsewhere [9], [5]. These results show that the MSE is a useful quality measure, provided that the nature of the error between the original and the regenerated signal is the same for all the speech samples compared. For another type of error the subjective degradation might be very different, even though the MSE is the same. In any delta modulation system, both quantization and slope overload contribute to the error. Usually these two degradations are difficult to analyze separately. Therefore, since a subjective test is always the ultimate quality measure if quality factors besides intelligibility are also tested, it is often convenient to compare the voice system under test with some standard signal that is easy to reproduce. N-bit logarithmic PCM as described by Smith [10] meets these requirements and was used in the present study as a reference signal for paired preference tests. Other standard quality rating methods are discussed in [11].

V. COMPUTER SIMULATION AND RESULTS

5.1 Preparation of the Data

Three sets of data were prepared, one for each of the sampling frequencies 8, 12, and 16 kHz. A set consists of the following two groups of two sentences:

"Joe took father's shoe bench out. Should we chase those young outlaw cowboys?",

and "We were away a year ago. May we all learn a yellow lion roar."

The four sentences have a total length of 10.75 sec, and are believed to be reasonably representative of conversational speech. The first group of two sentences contains most of the voiceless phonemes of English speech; whereas the second pair of sentences consists of voiced sounds only. A description of the sounds of English speech of General American Dialect, can be found in Flanagan [1], chapter 2.2. Note that the theory on which the coder is based neglects voiceless sounds. By choosing two groups of sentences, the expected differences in performance could be studied more conveniently.

All the sentences were spoken by a 35-year old male university professor with a western Canadian accent. The recordings were carried out in an anechoic chamber using an AKG D-200E dynamic microphone and a full-track Scully 280 tape recorder at a recording speed of 15 in/sec. Subsequently, the signal was lowpass filtered, sampled and digitized with an accuracy of 10 bits per sample and stored on digital magnetic tape. The analog tapes required for the listening tests were obtained by the reverse process.

5.2 Subjective Test Procedure

The tests were conducted in a quiet room. The same listeners, eleven male and one female university students, took part in all sessions, five had previous experience with listening tests. The test tapes were played back using the Scully 280 tape recorder and Sharpe HA-10-MK-II headphones, each equipped with an independent external volume control. Prior to a listening session, a few samples were played in order to permit the listeners to adjust their individual volume controls. Subsequently, no change in the loudness level was allowed.

The tests were based on preference only, and consisted of a series of pairs of utterances. For each pair of utterances presented, the subjects were asked to select the utterance they would prefer to listen to in a telephone conversation. The utterances of each pair were processed in different ways, but consisted of the same group of sentences. It is known that such tests exhibit a slight psychological bias for choosing the second utterance of a pair. For this reason the pairs were also played in reverse order.

Since each pair was played twice only, the total number of comparisons was too small to allow an arbitrary choice in case a listener was undecided as to which utterance to prefer. Therefore, six listeners were given the instruction to vote in favour of the first utterance in case they could not come to a decision. The other six were told to give their votes to the second utterance.

5.3 Results and Discussion

In this section the dependence of the S/Q ratio on the system parameters is studied. Only one parameter is varied at a time, all others are held at constant values. The standard set of parameters was:

Step multipliers	$P = 1.15$
	$Q = 1/P$
Direct step feedback coefficient	$\gamma = 0.2$
Maximum stepsize (quantizer level)	65 units
Minimum stepsize (quantizer level)	1 unit
Learning period NT	5 msec
Interval between re-calculation of the predictor parameters (frame length)	5 msec
Switch threshold; (zero-crossings)	15 per 5 msec
(signal energy)	15,000 units per 5 msec
Input signal amplitude range	-511 ... +512 units
RMS signal value for the first two sentences	66 units
RMS signal value for the second two sentences	81 units

5.3.1 Variation of the Step Multiplier P

Figures 5.1a, c, and d show the calculated S/Q ratio at different sampling rates. The values for P = 1.0 were obtained in simulations with a fixed stepsize that was optimized in terms of MSE. It is evident that the filter G added to the first predictor loop improved the performance. The gain is about 1 db for the sentences containing unvoiced sounds, and approximately 1.5 db for those consisting of voiced sounds only. In addition, a formal listening test

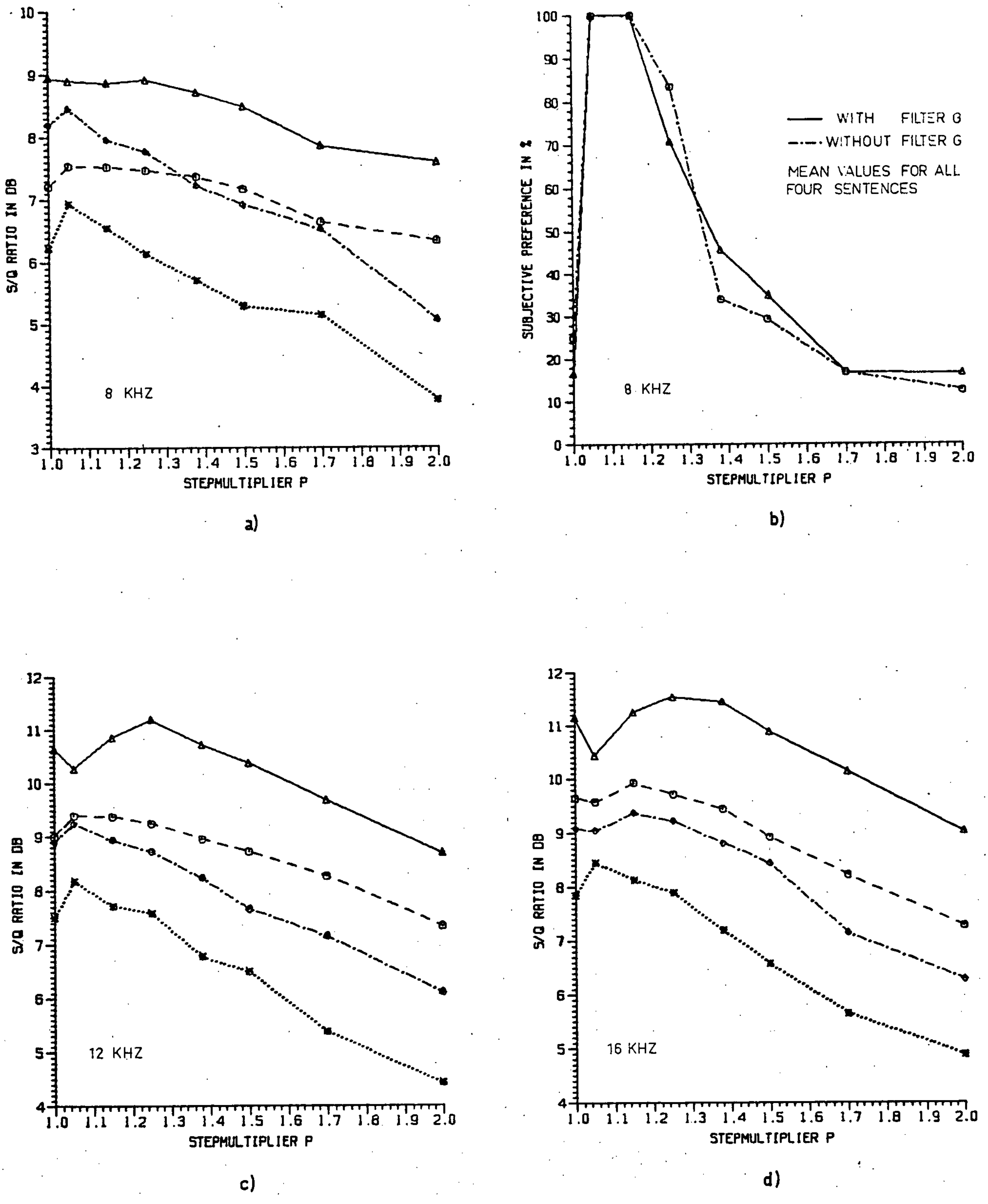


FIGURE 5.1 DEPENDENCE OF THE S/Q RATIO ON THE STEP MULTIPLIER P

"WE WERE AWAY A YEAR AGO. MAY WE ALL LEARN A YELLOW LION ROAR." } WITH FILTER G ——— WITHOUT FILTER G – – –

"JOE TOOK FATHER'S SHOE BENCH OUT. SHOULD WE CHASE THOSE YOUNG OUTLAW COWBOYS." } WITH FILTER G –·–· WITHOUT FILTER G ·········

confirmed a subjective improvement of the quality when the filter was used. The test was conducted for P = 1.15 and P = 1.05. All other parameters were set to their standard values. Between 69 and 71% of the listeners preferred the quality of the coder containing the filter G.

In another listening test the relation between the subjective quality and the step multiplier P was measured. The subjective preference curve, Figure 5.1b, was obtained by comparing sentences processed with various coefficients P, ranging from 1.0 to 2.0, to a reference sentence generated with P = 1.15. In a previous, informal listening test, the reference sentence was selected among all processed sentences as the one that yielded the best subjective quality. The ordinate in Figure 5.1b indicates the percentage of votes in favour of the quality of the speech encoder for a given value of the coefficient P. When comparing two reference sentences, each one should obtain the same number of votes (see last paragraph in section 5.2), corresponding to a subjective preference score of 100% in Figure 5.1b.

The calculated curves, Figures 5.1a, c, and d, do not show a very pronounced maximum. In particular, the S/Q ratio for a fixed step size is, on the average, about the same as for the best performance with a variable quantizer level. However, the subjective rating in Figure 5.1b shows a very distinct preference of step size multipliers P between 1.05 and 1.15 over all other values of P. In particular, the quality achieved with a constant step size is judged very poor, which is an expected result. The quantization error is due to either overload distortion or granular noise, depending on whether the step size is too small or too large relative to the signal being processed. The adaptation follows a certain time-invariant rule and will therefore yield smaller errors for small input signals, and larger errors for large signals. Thus, what can be gained is an increase in dynamic range rather than an inherent S/Q ratio advantage over a nonadaptive scheme. The adaptive system would be preferred by a listener because the error is signal amplitude dependent.

The preference curve Figure 5.1b was obtained in a formal listening test for a sampling rate of 8 kHz. An informal listening test using a sampling frequency of 12 kHz revealed the same preference of step multipliers P between 1.05 and 1.15. Therefore, Figure 5.1b is believed to be representative for several sampling rates.

The subjectively best performance of Jayant's system was obtained for P = 1.2 [5], which value coincides approximately with the values of P for best subjective quality of the present coder.

5.3.2 Variation of the Direct Feedback Coefficient γ

Figure 5.2a shows that no increase in the S/Q ratio is gained by adding the direct feedback of the step size to the second predictor loop. However, the preference curve in Figure 5.2b exhibits a distinct subjective preference for the quality achieved by all processors containing the direct feedback of the step size. Figure 5.2b was obtained in the same way as outlined for Figure 5.1b in section 5.3.1. With the exception of γ, which was set to 0.4, the reference sentence was generated using the standard set of parameters. The subjective preference is attributed to a reduction of crackling noise in the reconstructed speech.

It was also observed that the high-frequency content of the reproduced speech decreased as the feedback coefficient γ approached unity. If γ is set to 0.2 this lowpass effect is hardly noticeable, nevertheless, Figure 5.2b clearly shows a subjective preference for this value of γ over γ =0. For these two reasons, the standard set of parameters contains the value 0.2 for the step-feedback coefficient, even though the subjective optimum occurs for γ = 0.4.

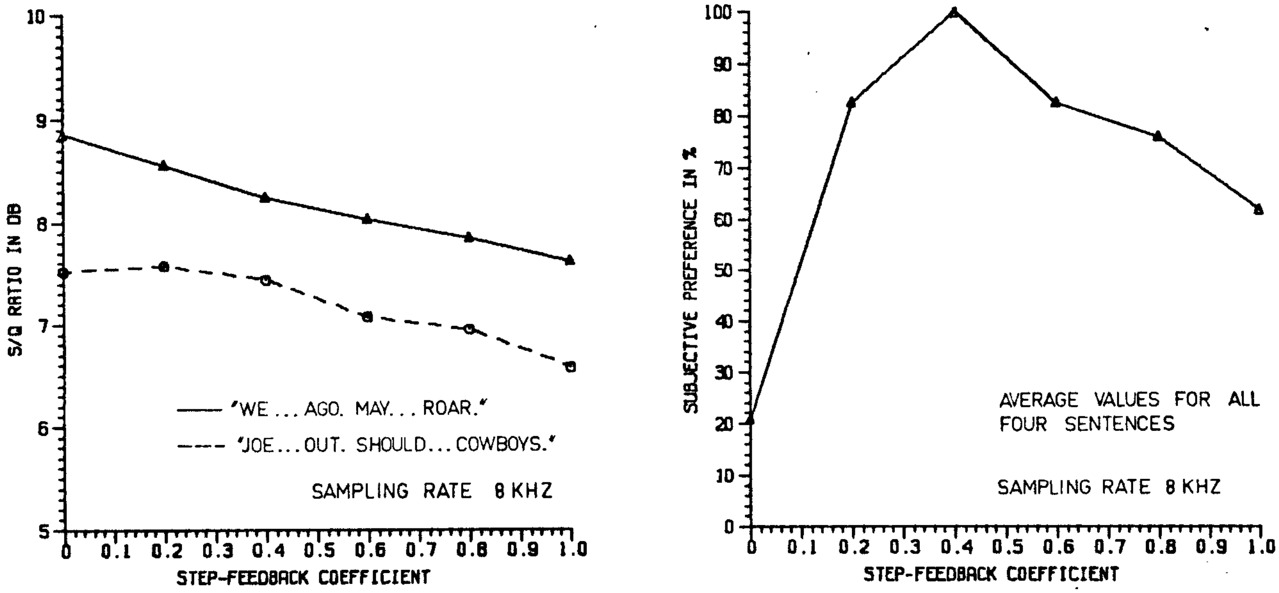


Figure 5.2 a) The S/Q ratio as a function of the coefficient γ b) Subjective preference as a function of the coefficient γ

5.3.3 Variation of the Frame Length

The interval between re-calculation of the predictor parameters is called the frame length of the coder. After each frame, the new predictor coefficients are transmitted to the receiver. The dependence of the S/Q ratio on the frame length is shown in Figure 5.3 for different sampling rates. As expected, the performance deteriorates when the readjustments occur less often.

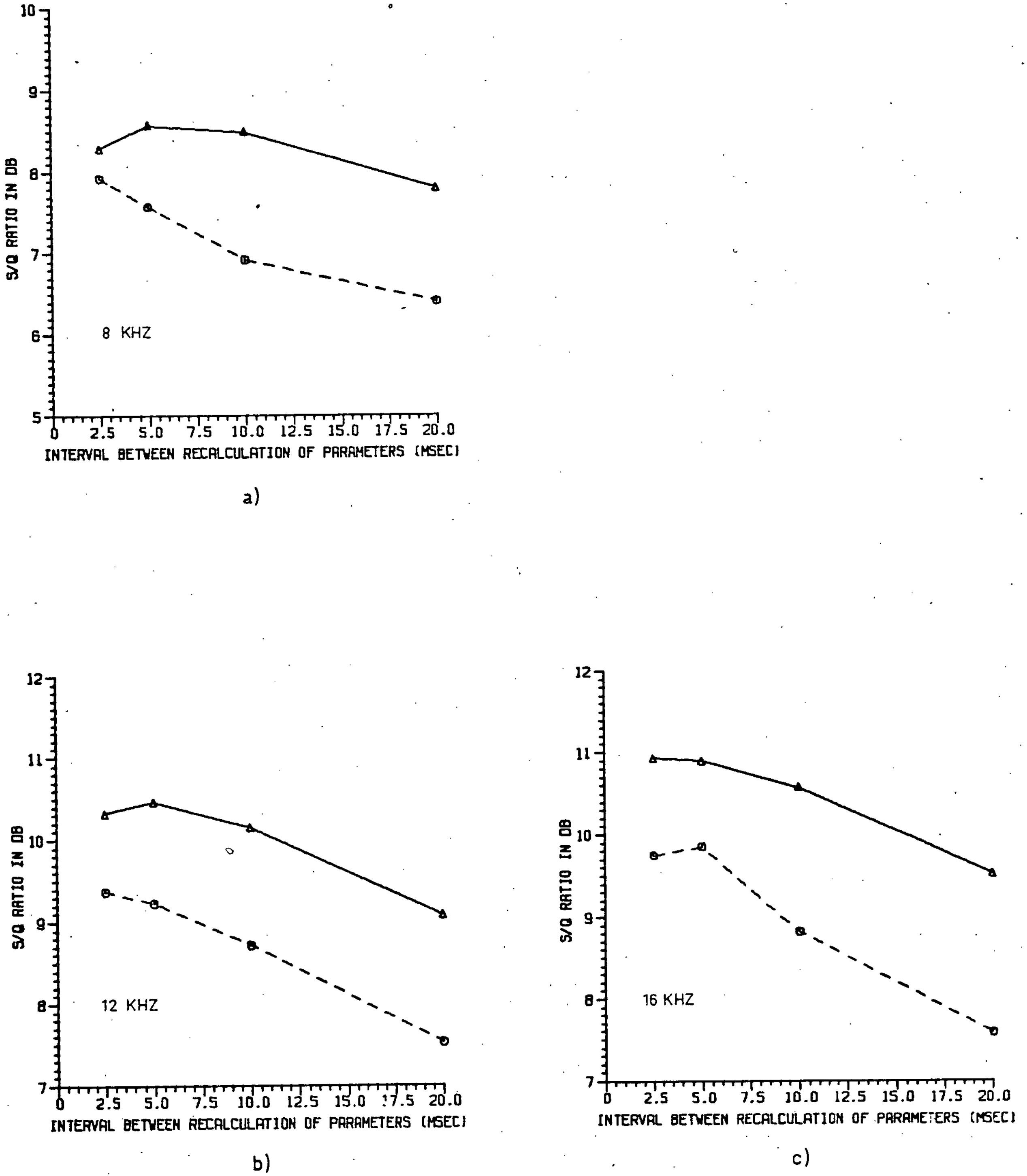


FIGURE 5.3 DEPENDENCE OF THE S/Q RATIO ON THE FRAME LENGTH

"WE WERE AWAY A YEAR AGO. MAY WE ALL LEARN A YELLOW LION ROAR." } ———

"JOE TOOK FATHER S SHOE BENCH OUT. SHOULD WE CHASE THOSE YOUNG OUTLAW COWBOYS." } – – – –

Individual phonemes have durations of about fifty to several hundred milliseconds. For most voiced sounds, the spectrum is nearly constant during this time; and the predictor coefficients need not be changed drastically. For other speech sounds, particularly during transitions from one phoneme to another, the spectrum may vary comparatively rapidly. For best performance, the coder must be able to adapt to these changes as quickly as possible. However, since the transitions are relatively short compared to the duration of voiced sounds, there is not much to be gained by readjusting the predictor in intervals considerably shorter than the average duration of a transition (10 - 20 msec). This effect is reflected in the curves for the sentences consisting of voiced sounds only, which show that there is only little or no improvement if the predictor is readjusted in intervals shorter than 5 msec. In the sentences containing voiceless sounds, a much larger number of periods with changing signal statistics occur. Therefore, these curves show a further improvement for frame lengths shorter than 5 msec.

5.3.4 Variation of the Learning Period

Figure 5.4 shows the dependence of the S/Q ratio on the length of the learning period for a sampling rate of 8 kHz. Curves for higher sampling frequencies were not measured, but are expected to look the same. A learning period of 40 input samples (5 msec) seems to be sufficient to determine the short-term signal statistics with adequate accuracy. This corresponds to results previously obtained by Davisson [13].

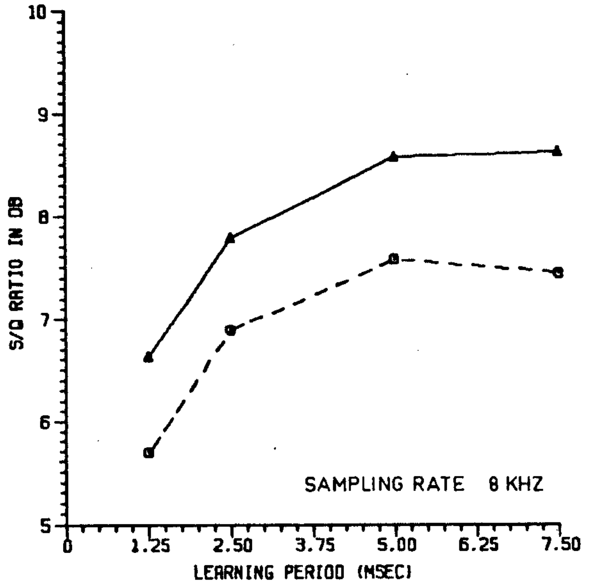


Figure 5.4 The S/Q ratio as a function of the learning period for a sampling frequency of 8 kHz

5.3.5 Variation of the Switch Threshold

Figures 5.5a and 5.5b indicate that the variation of the threshold for the switch in the first predictor loop (see Figure 3.1) has very little influence on the overall S/Q ratio. However, it was found in some informal listening tests, that introducing this switch caused disappearance of some of the noise during quiet input periods and some of the distortion at the beginning of words, especially after such quiet periods.

Figure 5.5a shows the S/Q ratio as a function of the number of zero-crossings per 5 msec which causes the first loop to be disabled by opening the switch. If the threshold is set to zero, the pitch prediction is never in operation. Thus, the 1.5 db increase of the S/Q ratio between the switch thresholds of 0 and 10 zero-crossings per 5 msec reflects the improvement achieved by the first predictor loop.

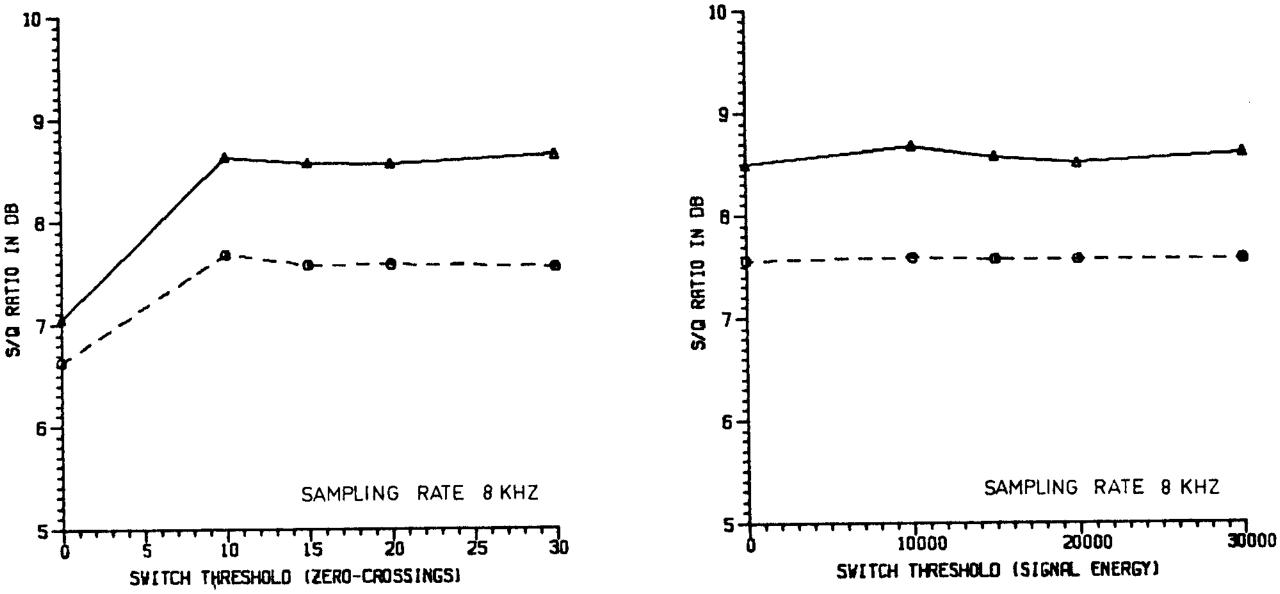


Figure 5.5 a) Dependence of the S/Q ratio on the switch threshold set by the number of zero-crossings

b) Dependence of the S/Q ratio on the switch threshold set by the signal energy

Figure 5.5b shows the S/Q ratio versus the switch threshold which is dependent on the signal energy. A number, corresponding to the signal energy, is calculated by summation of the squares of the signal samples during the last learning period. If this sum is less than the threshold, the switch in the pitch predictor is opened.

Both of the above curves were also calculated for sampling frequencies of 12 and 16 kHz, but are not shown because they were very similar to the above.

5.3.6 Variation of the Product of the Step Multipliers P and Q

In Figure 5.6 the dependence of the S/Q ratio versus the product of the quantizer level multipliers P and Q is shown. The maximum occurs for values of P·Q between 1.0 and 1.05.

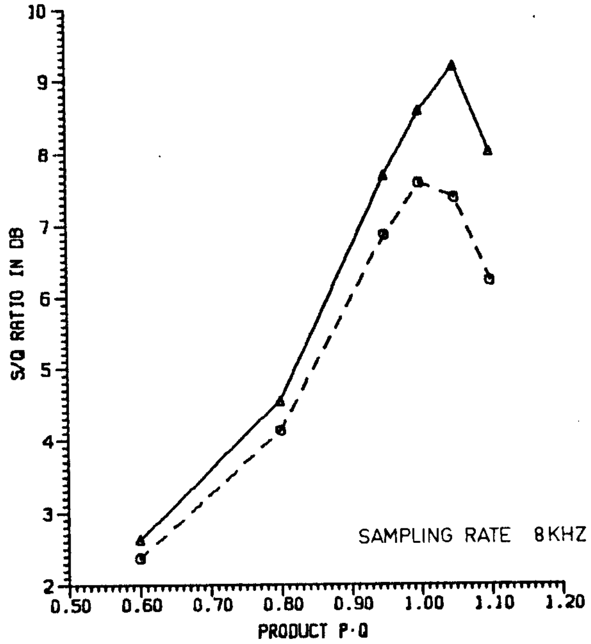


Figure 5.6 S/Q ratio as a function of the product P·Q

Jayant's bound, which is

$$(P \cdot Q)_{opt} = 1, \qquad (2.1)$$

(see section 2.1) cannot be applied exactly, because the assumption that step size adaptations using the multipliers P and Q are equally probable, is not satisfied for the present coder. The upper and lower limits imposed on the quantizer level affect the probabilities for P-type and Q-type adaptations. Also, the direct step-feedback in the second loop tends to make the prediction of D_n exceed the true value, and thus favours the use of the multiplier Q. Nevertheless, the requirement that the step size should not tend to increase beyond limits or to decay to zero, still applies. Jayant considers the ratio R(N) of the magnitudes of $L_{T'+N}$, the step size at the sampling instant T'+N, and $L_{T'}$, the step size at the sampling instant T'. Denoting the number of P-type and Q-type adaptations in the interval N by Np_o and Nq_o respectively, Jayant writes [4]

$$R(N) = \frac{|L_{T'+N}|}{|L_{T'}|} = P^{Np_o}.Q^{Nq_o} = (P^{p_o}.Q^{q_o})^N$$

For N→∞ p_o and q_o tend to the probabilities p_{opt} and q_{opt} respectively, for an optimum adaptation. Thus,

$$\lim_{N\to\infty} R_{opt}(N) = \lim_{N\to\infty} (P^{p_{opt}}.Q^{q_{opt}})^N .$$

It was mentioned above that the step size must not show a tendency to increase beyond limits or decay to zero. Therefore, the asymptotic ratio $R_{opt}(\infty)$ must be finite and non-zero. A necessary and sufficient condition is

$$P^{p_{opt}}.Q^{q_{opt}} = 1 . \qquad (5.1)$$

Experimental results confirmed this requirement. The values p_o and q_o for the two sentences containing only voiced sounds were 0.400 and 0.600 respectively. The optimum predictor performance, in terms of S/Q ratio, was achieved for

$$P\cdot Q = 1.05,$$

where P = 1.15 and Q = 0.913043 (see Figure 5.6).

Application of the above condition yields

$$1.15^{0.4}\cdot 0.913043^{0.6} = 1.00129,$$

which satisfies (5.1).

5.3.7 Comparison with Logarithmic PCM

The same four test sentences were encoded using 3- , 4-, and 5-bit log PCM with a sampling rate of 8 kHz. They were then compared in formal listening tests to the quality achieved by the adaptive predictive delta modulator (APDM).

The compression characteristic for a log PCM quantizer is defined by

$$y = \frac{V \cdot \log(1 + \mu x / V)}{\log(1 + \mu)} \cdot \mathrm{sgn}(x),$$

where y represents the output voltage corresponding to an input signal voltage x, μ is a dimensionless parameter which determines the degree of compression, and V is the compressor overload voltage [10]. Since the four test sentences were digitized with an accuracy of 10 bits per sample, V was set to 512. The parameter μ was chosen equal to 100 to make the listening tests compatible with most other such tests in the literature.

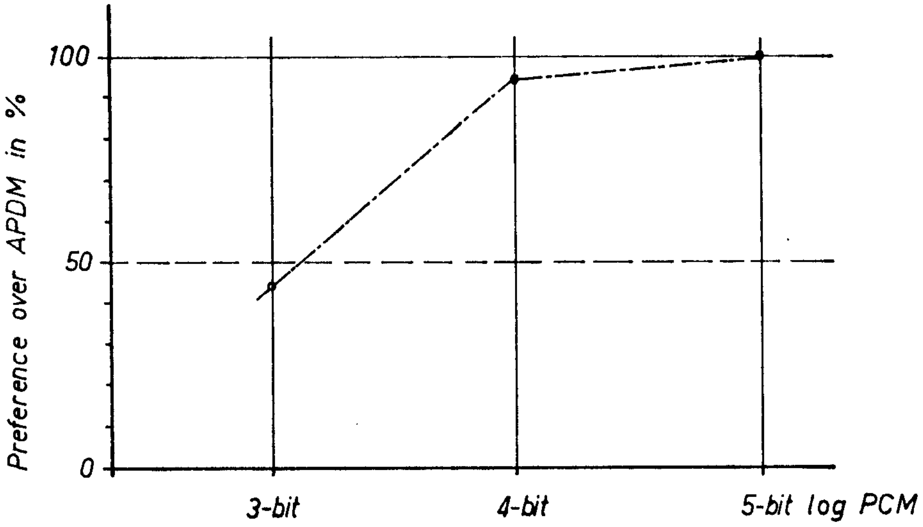


Figure 5.7 Comparison of the Adaptive Predictive Delta Modulator to logarithmic PCM

The results of the subjective tests (see Figure 5.7) show that the quality of the speech processed by the APDM was slightly better than that of 3-bit log PCM. The crossover point occurs at 3.1 bits per sample. Even though the granular noise of the APDM system compares favourably to the granular noise of 4-bit log PCM, many listeners still preferred the quality of 3-bit log PCM to the quality of the APDM coder because of the greater high-frequency content of 3-bit log PCM. Also, 3-bit log PCM reproduces voiceless sounds better despite its very coarse quantization.

VI. ADDITIONAL RESULTS AND RECOMMENDATIONS

This chapter contains information concerning additional experiments and observations, as well as some recommendations for further research.

6.1 Improvement of the Formant Redundancy Reduction

Increasing the number of previous samples used in the formant predictor should result in an improvement of the coder performance. A formant predictor based on eight previous samples was implemented on the computer, but it turned out to be unstable. The poles of the filter transfer function were restricted in the same way as outlined in section 3.2, but, this did not solve the problem completely. In addition, this operation requires a fair amount of computation to find the poles of the filter transfer function. A better and simpler solution in terms of computation is indicated in Haskew et.al. [14]. Instead of finding the poles of the filter transfer function, Haskew et.al. [14] calculated the area function of an acoustic tube which corresponds to the filter defined by the predictor coefficients. Unstable predictor coefficients result in an area function with some negative values. Thus, such instabilities can be detected easily.

6.2 Improvement of the Reproduction of Voiceless Sounds

The adaptive predictive delta modulator is based on theories that consider only voiced parts of speech. Furthermore, the spectrum of a voice signal falls off at 6 to 12 db per octave at frequencies above 500 Hz. These are the two main factors contributing to the poor reproduction of voiceless sounds.

In order to enhance the high-frequency content of the reproduced signal, the input signal spectrum was shaped using a digital filter. The pre-emphasis implemented began at 600 Hz and increased linearly with frequency to +12 db at 3000 Hz. It was found that for the reasons explained in section 3.1, the second predictor loop of the encoder was not able to handle the additional high-frequency content of the input signal. The delay contained in this part of the coder represents a severe problem when the input signal amplitude is fluctuating rapidly. These difficulties could probably be circumvented by increasing the number of quantizer levels. This would allow the instantaneous selection of several step sizes, thus partially bypassing the delay problem. Extensive research concerning coders using adaptive multi-level quantizers has been carried out by Cummiskey [12]. Another approach is described in [14], where vocoder techniques are used to transmit voiceless sounds.

Unfortunately, improvements mentioned above may require that the bit-rate be increased.

6.3 Effect of Channel Noise

The effect of transmission errors due to channel noise was not studied because the digital channels presently used commercially are virtually error-free. The error rates are of the order of one bit per 10^6 bits. It is known that the performance of Jayant's scheme degrades rapidly when the error probability becomes considerably larger than 10^{-6} [15]. It is expected that the APDM coder will show a similar behaviour.

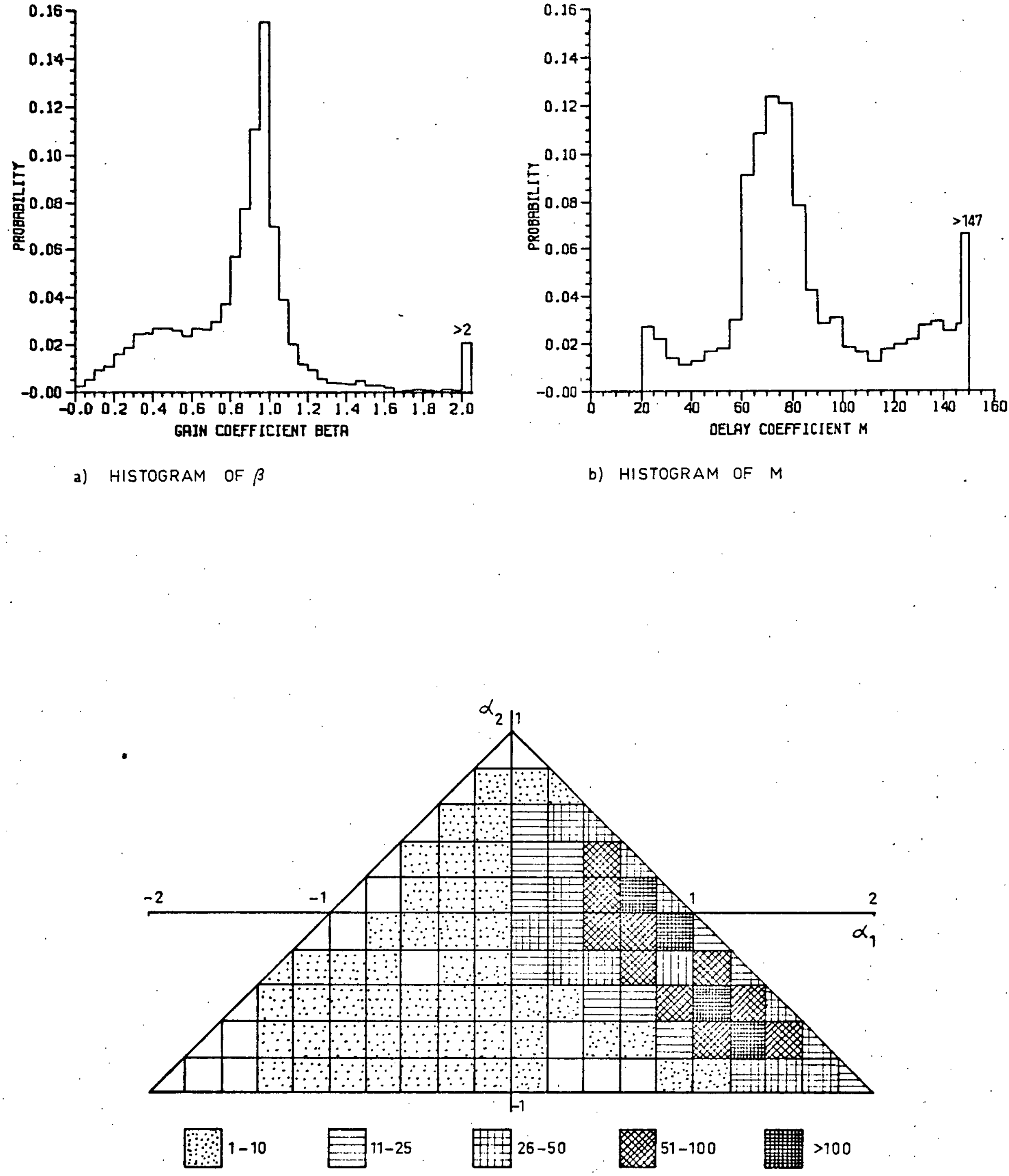
6.4 Quantization of the Predictor Parameters

Figures 6.1a, b, and c show histograms of the gain parameter β, the delay coefficient M, and the two predictor coefficients α_1 and α_2. All three graphs exhibit a clustering of the parameter values. Therefore, to minimize the number of bits required to transmit the coder parameters, nonlinear quantization or variable length coding should be used. Note, however, that the delay coefficient M takes on integer values only and may not be quantized any further. Variable length encoding is possible but requires additional buffer storage, thus increasing the coder complexity.

Table 6.1 gives an estimate of the number of bits required to transmit the predictor parameters. The delay parameter is given a range of 20 to 146, thus requiring 7 bits for its encoding. The largest number, 146, is reserved for transmitting the decision voiced/unvoiced. Parameter M becomes irrelevant when the decision requires disabling of the first predictor loop by opening the switch. In this case, the transmitter sets M equal to 146, which tells the receiver to switch the pitch prediction off. This procedure saves one bit that would otherwise be necessary to transmit the switch operation separately. The estimate of the bit requirement for the gain coefficient β is based on results of a nonlinear quantization of this parameter in Kelly et al. [16].

Table 6.1

Parameter	Bit Requirement per frame
M	7
β	3
α_1	5
α_2	4
total	19 bits/frame



a) HISTOGRAM OF β

b) HISTOGRAM OF M

c) HISTOGRAM OF α_1 AND α_2, SHOWING THE NUMBER OF CALCULATED VALUES FALLING WITHIN A CERTAIN RANGE

FIGURE 6.1 HISTOGRAMS OF THE CODER PARAMETERS, CALCULATED FOR ALL FOUR SENTENCES

An experimental quantization of the coefficients α_1 and α_2 to 5 and 4 bits respectively, reduced the S/Q ratio of the two sentences containing no voiceless sounds by 0.65 db. The corresponding degradation for the other two sentences was 0.8 db. It is expected that a proper nonlinear quantization, using the same number of bits, would yield considerably better results.

Table 6.1 indicates that at 200 frames per second 3800 bits are required to transmit the predictor coefficients. The overall bit-rate will therefore reach 11.8 kbits/sec when the input signal is sampled at 8 kHz.

VII. SUMMARY AND CONCLUSIONS

A new adaptive predictive delta modulator was simulated on a digital computer and optimized for speech signals. The main objective was to keep the transmission bit-rate as low as possible. Therefore, an efficient redundancy reduction technique had to be applied. Speech signals contain segments with low correlation between the samples, such as voiceless sounds, transition periods between two sounds, and pauses; but the signal energy is mainly concentrated in voiced segments with high correlation between the samples. Redundancy reduction methods therefore concentrate on these segments. The main sources of redundancy in such parts of the speech signal are peaks in the short-term spectral envelope, also called formants, and the pitch, that is, a quasi-periodicity of the signal.

The first stage of the coder calculates the difference between the true signal value and a prediction, derived from the value one pitch period before, thus reducing the pitch redundancy. Subsequently, the difference signal is filtered in the second stage of the coder, where a recursive filter removes some formant redundancy. The remaining, now less correlated error signal is quantized to two levels and transmitted.

Since speech signals lack periodicity in voiceless sounds and during quiet periods, the pitch predictor contains a switch, activated by a threshold dependent on the zero-crossing rate and the signal energy. This addition resulted in a subjectively better performance during quiet signal periods and at the beginning of words. An improvement of 1 to 2 db of the S/Q ratio was achieved by a filter inserted into the pitch predictor to eliminate some quantization noise fed back to this part of the coder.

For the pitch predictor a delay and a gain coefficient are all that are required. The number of parameters used in the second stage depends on the number of formants to be removed. The coder described here reduces redundancy due to the first formant only. Two parameters are required for this purpose. Assuming that the signal statistics are stationary over short periods of time, all four parameters are recalculated in intervals of 5 msec (frame length of the coder) and optimized in terms of mean square error over the last 5 msec (learning period of the coder). Measurements of the S/Q ratio for different lengths of the learning period showed that 5 msec are sufficient to determine the signal statistics with adequate accuracy. Similar calculations for the frame length indicated that no substantial improvement of the S/Q ratio was achieved with frame lengths shorter than 5 msec.

Contrary to the four coder parameters that are readjusted only every 5 msec, the quantizer level changes each time a new input sample has been processed. The adaptation rule is simple; according to whether the new error signal has the same sign as the previous one, the quantizer level is multiplied by a constant factor greater or less than one. During periods of slope overload the quantizer level increases exponentially, and when the signal energy drops to low values the quantizer level decreases accordingly. Thus, for any signal level, the quantizer adjusts its level to the optimum value. The computer simulations of the system revealed some difficulties originating from this exponential self-adaptation. Due to the delay of one sampling period, the adaptation was sometimes not able to follow fast fluctuating input signals. The adaptation was disturbed, causing instabilities and poor reproduction of the high-frequency content of such signals. This effect was observed in particular for voiceless sounds. Since the adaptation scheme is non-linear and has memory, it becomes mathematically intractable. Experimentally no completely satisfactory solution of the stability problem was found. However, limiting the maximum allowable quantizer level to one eighth of the peak signal amplitude and adding a direct feedback of the quantizer level to the second stage of the coder improved the stability considerably. Occasionally, some instabilities still occur. They result in clicks in the reproduced speech signal.

The coder was simulated on a digital computer and optimized for sampling rates of 8, 12, and 16 kHz, using objective calculations of the S/Q ratio as well as subjective listening tests. In some cases where the measured S/Q ratios were almost identical and did not allow any distinction in performance, the subjective tests exhibited strong preferences. This observation enhances the necessity to evaluate the performance of such systems by subjective methods in addition to the objective calculation of the mean-square errors.

Comparisons with speech of the same bandwidth that was encoded by a 3-bit and a 4-bit log PCM system with the same sampling rate of 8 kHz, showed that at this sampling rate the subjective quality of the adaptive predictive delta modulator was equivalent to the subjective quality of 3.1-bit log PCM. Since the granular noise of the adaptive coder is similar to that of 4-bit log PCM, it is expected that if the reproduction of high frequencies could be improved, the adaptive coder would compare favourably to 4-bit log PCM.

The quantization and encoding of the four coder parameters that have to be transmitted to the receiver was not studied in detail.

However, preliminary results show that when the voice signal is sampled at 8 kHz, an additional 3800 bits/sec are required to transmit the parameters. This suggests a final transmission bit-rate of 11.8 kbits/sec. A PCM signal with the equivalent subjective quality requires 24 kbits/sec. Thus, a data compression of 2:1 has been achieved.

One could probably reduce the required bit-rate from 11.8 kbits/sec to 9.6 kbits/sec (a standard rate for telephone traffic) without seriously degrading speech quality. Reduction of the sampling rate from 8 kHz to 7 kHz should not affect the quality significantly, provided the input speech bandwith is reduced from 3.75 kHz to 3.4 kHz [17]. Increasing the frame length from 5.0 msec to 7.5 msec should reduce the S/Q ratio by less than 1/2 db (see Figure 5.3). Finally, additional work to improve high frequency fidelity might well result in a 9.6 kbits/sec coder whose quality rivals that of 4-bit log PCM.

REFERENCES

1. Flanagan, J.L., "Speech Analysis, Synthesis, and Perception", Second Edition, Springer, 1972.

2. Bayless, J.W., Campanella, S.J., and Goldberg, A.J., "Voice Signals: bit-by-bit", IEEE Spectrum, pp. 28-34, Oct. 1973.

3. Atal, B.S., and Hanauer, S.L., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", JASA, vol. 50, pp. 637-655, 1971.

4. Jayant, N.S., "An Adaptive Delta Modulation with a One-Bit Memory", BSTJ, vol. 49, no. 3, pp. 321-342, March 1970.

5. Jayant, N.S., and Rabiner, A.E., "The Preference of Slope Overload to Granularity in the Delta Modulation of Speech", BSTJ, vol. 50, no. 10, pp. 3117-3125, Dec. 1971.

6. Atal, B.S., and Schroeder, M.R., "Adaptive Predictive Coding of Speech Signals", BSTJ, vol. 49, no. 65, pp. 1973-1986, Oct. 1970.

7. Schroeder, M.R., "Vocoders: Analysis and Synthesis of Speech, Proc. IEEE, vol. 54, no. 5, pp. 720-734, May 1966.

8. Faddeev, D.K., and Faddeeva, V.N., "Computational Methods of Linear Algebra", English Translation by R.C. Williams, San Francisco: W.H. Freeman, 1963, pp. 144-147.

9. Levitt, H., McGonegal, C.A., and Cherry, L.L., "Perception of Slope-Overload Distortion in Delta-Modulated Speech Signals", IEEE Trans. Audio Electroac., vol. AU-18, no. 3, pp. 240-247, Sept. 1970.

10. Smith, B., "Instantaneous Companding of Quantized Signals", BSTJ, vol. 36, pp. 653-709, May 1957.

11. "IEEE Recommended Practice for Speech Quality Measurements", IEEE Trans. Audio Electroac., vol. AU-17, pp. 227-246, Sept. 1969.

12. Cummiskey, P., "Adaptive Differential PCM for Speech Processing", Newark College of Engineering, D. Eng. Sc. Thesis, 1973.

13. Davisson, L.D., "Theory of Adaptive Data Compression", Recent Advances in Communication Systems, vol. 2, A.V. Balakrishnan Ed., New York Academic Press 1966.

14. Haskew, J.R., Kelly, J.M., Kelly, R.M., and McKinney, T.H., "Results of a study of the Linear Prediction Vocoder", IEEE Trans. Commun. Technol., vol. COM-21, no. 9, pp. 1008-1014, Sept. 1973.

15. Chang, K.Y., and Donaldson, R.W., unpublished work.

16. Kelly, J.M., et al., Final report on predictive coding of speech signals, contract no. DAAB03-69-C-0338, Bell Laboratories, June 1970.

17. Donaldson, R.W., and Chan, D., "Analysis and Subjective Evaluation of Differential Pulse-Code Modulation Voice Communication System", IEEE Trans. Commun. Technol., vol. COM-17, pp. 10-19, Feb. 1969.