ADAPTIVE TRANSFORM CODING OF SPEECH SIGNALS

RICHARD JAMES PINNELL

B. Eng., McGill University
(1980)

A Thesis Submitted to the
Faculty of Graduate Studies and Research
in Partial Fulfillment of the Requirements for the
Degree of

MASTER OF ELECTRICAL ENGINEERING

at

McGill University
Montreal, Canada

ACKNOWLEDGEMENTS

I would like to thank my thesis supervisor, Dr. P. Kabal for his valuable encouragement and guidance in both the experimental aspect of this work and in the preparation of this thesis.

LIST OF FIGURES

LIST OF FIGURES

ABSTRACT

Frequency domain coding techniques have recently received considerable attention. Prominent among these techniques, adaptive transform coding offers excellent speech quality for low to medium data rates (8-16 kb/sec). Adaptive transform coders divide speech into frequency components by using a suitable transform and transmit these components using pulse code modulation (PCM). Three basic issues in the design of adaptive transform coders are:

(1) Selection of the best transform

(2) Selection of the best quantization strategy

(3) Selection of a spectral parameterization technique

This thesis discusses design considerations with emphasis on finding variants of adaptive transform algorithms amenable to hardware implementation. In this context coder performance using reduced frame lengths is presented. Objective and subjective performance reduction, caused by frame boundary discontinuities and low-pass filtering effects are investigated as the primary sources of perceptual distortion. Results from two computer simulations of adaptive transform coders using all-pole and homomorphic spectral fits are presented.

SOMMAIRE

Les techniques de codage dans le domaine fréquentiel ont récemment fait l'objet d'une attention considérable. Le codage de transformées par adaptation y occupe une place de choix parce qu'il permet une excellente qualité de transmission de la parole pour des débits faibles ou moyens (8-16 kHz). Les systèmes de codage de transformées par adaptation effectuent une segmentation de la parole en diverses composantes fréquentielles grâce à l'utilisation d'une transformée appropriée et transmettent ces composantes à l'aide de la modulation par impulsion et codage (MIC). Les codeurs de transformées par adaptation sont associés à trois questions fondamentales:

(1) Sélection de la meilleure transformée
(2) Sélection de la meilleure stratégie de quantification
(3) Sélection d'une technique de définition des paramètres spectraux

La présente thèse traite de considération théoriques et met l'accent sur la détermination de variantes d'algorithmes relatifs aux transformées par adaptation, pouvant être traduits en systèmes mécaniques. Dans ce contexte, on présente les performances de codage, faisant appel à des longueurs de trames réduites. On analyse les réductions des performances objectives et subjectives résultant des discontinuités des limites de trames et des effets de filtrage passe-bas, envisagées comme les sources principales de la distorsoin liée à la perception. On examine enfin les résultats de deux simulations par ordinateur de codeurs de transformées par adaptation, faisant appel à des courbes homomorphiques spectrales et entièrement polaires.

CHAPTER 1

INTRODUCTION

The objective of speech coding is to transmit the highest quality speech over the least possible channel capacity while employing the least complex coder. Coder efficiency in channel utilisation is, however directly linked to coder complexity and cost. Fortunately, advances in LSI (large scale integration) technology are now making available more sophisticated digital signal processing devices at reduced costs. Thus, telephone networks are moving toward digital switching and processing of voice signals. Investigations of more complex coding schemes are continuing in the light of these recent LSI technology advances. This new technology offers greater system flexibility and considerable cost advantage.

Speech coders can be divided into two distinct classes; waveform coders and source coders (vocoders). Waveform coders strive for facsimile reproduction of the signal waveform. By observing the statistics of a signal, the waveform coder can be tailored to the signal resulting in reduced coding error, and a more signal specific coder. Source coders employ a minimal parametric description derived from a hypothesis of speech production. Consequently, these units can

be operated at lower transmission rates. Source coders are also more sensitive to speaker variation and background noise than are those of the waveform classification.

In speech coding, transmission rates determine which class of coders is the more effective. Above 5 kb/sec waveform coders offer communication and toll quality speech. Speech quality for waveform coders declines very rapidly below this figure. At lower rates (below 5 kb/sec.) source coders can be used, to produce synthetic quality speech [1].

Waveform coding can be performed in either time or frequency domains. Two examples of the latter are subband and adaptive transform coders. Frequency domain coding is accomplished by dividing speech into a number of frequency bands by using a filter bank, or into frequency components by using a block transformation. These frequency components are then quantized and encoded. A replica of the input waveform can be re-synthesized by decoding the frequency components and subsequent filter bank summation or, inverse transformation if a transform was originally used. Both methods assume the input signal is quasi-stationary and can be locally modelled by a short time spectrum. Perceptually important components of the short time spectrum must be isolated and transmitted without incurring excessive delay or distortion.

Additional demands are placed on speech coding schemes by the context in which they are used. A likely area of application for speech coders is in telephony. Since a telecommunications carrier has little control over the type of signals the network will support, it is highly desirable that speech coders support a variety of input

signals including modem signals. In a military context encryption is made possible by the digital nature of speech coders. Since good speech quality is not essential, maximum speech compression is one of the primary objectives.

The mathematical principles behind transform coding were first formulated by Huang in a paper entitled "Block Quantization of Correlated Gaussian Random Variables" [2]. Huang develops a procedure for quantizing blocks of correlated Gaussian random variables. A linear transformation first converts the dependent random variables into independent random variables. Then the transformed random variables are efficiently quantized one-by-one until the bits allocated for the block are exhausted. A second linear transformation constructs (from the quantized values) the best estimate of the original variables in a mean square error sense. Huang develops the best choice for each transform and an approximate expression is derived for the number of bits assigned to each of the quantized variables. Segall [3] in a paper entitled "Bit Allocation and Encoding for Vector Sources" obtained a more precise expression for the allocation of available bits to quantization of the transformed variables.

Zelinski and Noll [4] developed a speech coder based on the principles discussed by Huang and Segall. Their important contribution was an adaptive quantization strategy employing the discrete cosine transform. The adaptation is controlled by a short term spectrum obtained from the transform coefficients prior to quantization. The short term spectrum is then parameterized and sent to the receiver as side information. A second paper by Zelinski and

Noll [5] presents refinements to the side information parameterization technique. The paper discusses improvements to the quantization strategy aimed at improving the subjective performance of the coder. Two papers by Tribolet and Crochiere [6,7] discuss adaptive transform coders which employ all-pole modelling of the short term spectrum. The papers compare sub-band coders and adaptive transform coders in the context of an analysis/synthesis framework. Cox and Crochiere [8] in a paper entitled "Real-Time Simulation of Adaptive Transform Coding" develop a homomorphic model for parameterization of the short term spectrum. Cox claims the technique performs as well as the all-pole model and is easier to implement in a real time coding context. Numerous authors have contributed to the development of transform coding directly and indirectly but the above papers are the most often quoted as references.

This thesis reviews current transform coding strategies and is directed towards finding variants of adaptive transform algorithms amenable to hardware implementation. Chapter 2 presents the theory and basic structure of transform coding. The applicability of various transforms and bit assignment algorithms is discussed from a theoretical standpoint for their usefulness in coding speech. Chapter 3 considers various adaptive transform coding strategies employing a short term spectrum to adapt the transform coefficient quantizers. Three techniques of parameterizing the short term spectrum for transmission to the receiver are presented. In Chapter 4 the results of listening tests are used to evaluate coded speech generated by computer simulations. These coders use either all-pole or homomorphic modelling of the short term spectrum. Impairments in speech quality and their causes are identified. Techniques to combat these

impairments are implemented. The effect of reduced frame sizes and perceptual observations on frame boundary discontinuities are discussed. Quantization noise shaping and noise insertion into low energy frequency bands are studied. Chapter 5 summarizes the important results and presents conclusions based on results obtained from the simulation of these coders.

CHAPTER 2

THE THEORY OF TRANSFORM CODING

In this section the theory of transform coding is developed and related topics are discussed. The treatment emphasizes important elements including the basic structure of transform coding, selection and justification of the mean square error distortion criteria, and discussion of an optimal transform and bit assignment rule. Sub-optimal transforms are introduced and compared with the discrete cosine transform. The latter is known to be a good choice and resistant to frame discontinuity distortion. The following presentation includes sufficient theory to support the subject matter.

2.1 BASIC TRANSFORM CODING

Mathematical concepts of transform coders are depicted in Figure 2-1. A frame buffer arranges N succesive source samples x(n) into the source vector $\underline{X}$. The speech is assumed to be bandlimited with the sampler satisfying the sampling theorm in order to avoid aliasing. A linear transformation is performed on the source vector $\underline{X}$ to obtain the transform coefficient vector $\underline{Y}$. Such an operation can be represented by the matrix equation 2.1 where A is unitary.

eq. 2.1 $\underline{Y} = A \cdot \underline{X}$

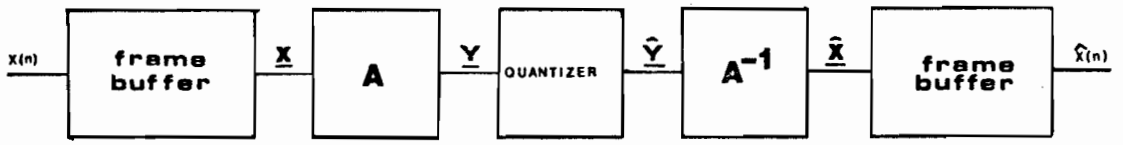
Reconstructed output samples are obtained from the quantized transform vector $\hat{\underline{Y}}$ by inverse transformation. The matrix equation representing this operation is

eq. 2.2 $\hat{\underline{X}} = A^{-1} \cdot \hat{\underline{Y}}$

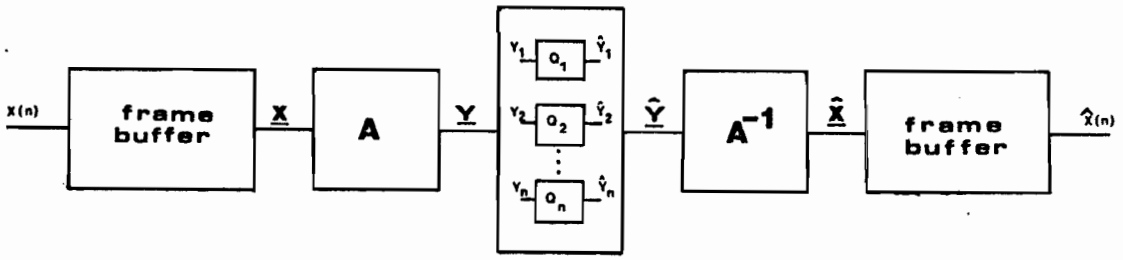
The overall mean squared overall distortion of the coding scheme is equal to the total quantization error i.e.

eq. 2.3 $1/N \; E\{ (\underline{X}-\hat{\underline{X}})^T \cdot (\underline{X}-\hat{\underline{X}}) \} = 1/N \; E\{ (\underline{Y}-\hat{\underline{Y}})^T \cdot (\underline{Y}-\hat{\underline{Y}}) \}$

Minimization of distortion requires an appropriate quantization strategy and transform. As will be shown later a necessary condition for minimum coding error is that every transform coefficient suffer the same amount of distortion.



A GENERAL TRANSFORM CODING



B BASIS RESTRICTED TRANSFORM CODING

TRANSFORM CODING BASICS

FIGURE 2-1

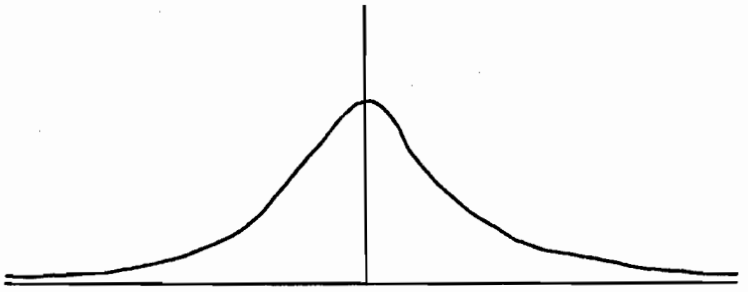
2.2 QUANTIZATION STRATEGY

Quantization strategy refers to the technique employed to quantize the transform coefficient. Basis restricted transform coding schemes quantize the coefficients y_i i=1,2,3...n independently. Only basis restricted quantization schemes are considered here. This can be rationalized by considering the source vector $\underline{X}$ to be an N dimensional Gaussian random variable with zero mean. A nonsingualar matrix A operates on $\underline{X}$ to yield the transform vector $\underline{Y}$ of uncorrelated random variables. Since $\underline{X}$ is Gaussian, $\underline{Y}$ is Gaussian and its components y_i are not only uncorrelated but actually independent. Huang [2] shows that the basis restricted quantization schemes are optimal when the transform coefficients are independent.

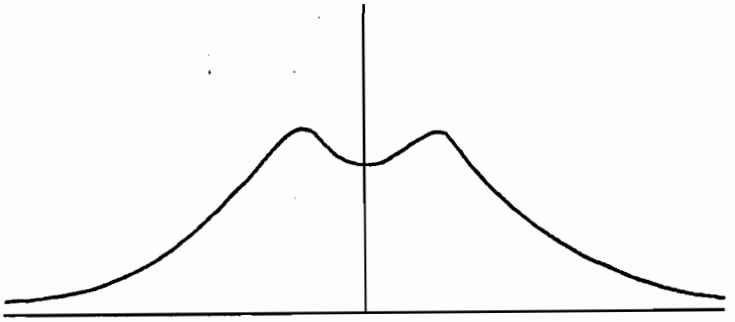
The quantization strategy is characterized by step size adaptation and bit assignment rules. Overall distortion given in equation 2.3 can be reduced by assigning quantizers with suitable (generally different) number of levels to each of the N transform coefficients. The distribution of the number of quantization levels for each quantizer is known as the bit assignment. The step size adaptation of the quantizers is accomplished by pre-scaling the transform coefficients by an estimate of the coefficients. The distribution of the coefficients, in general, depends on the transform and source signal. Analytic calculation of the distribution function is too difficult to yield meaningful results except in special cases. Goldberg and Cosell [9] obtained numerical results for discrete cosine transform (DCT) coefficients showing the coefficient distribution to lie between the Gaussian and Laplace distribution. However, the effect of pre-scaling transform coefficients is to make the

distribution bi-modal about +1 and -1. In the limit of perfect energy estimation, the distribution approaches a pair of impulses at +1 and -1. The pre-scaling effect is illustrated in Figure 2-2 and is valid for any distribution. In any event, the distribution of the scaled transform coefficients is such that an improvement over single variable time domain quantization (PCM) can be expected. This is explained in Section 2.5 by the concept of transform coding gain.

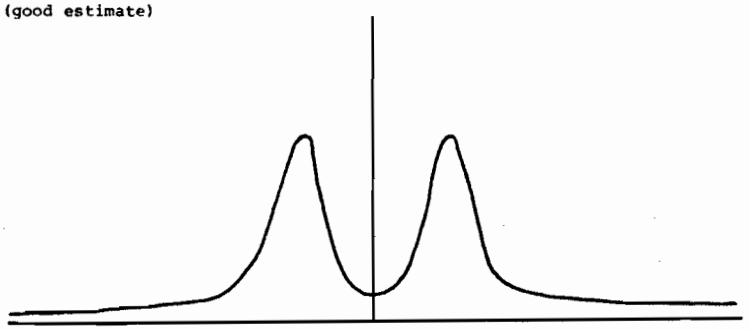
In order to minimize the distortion measure both the transform and the quantization bit assignment must be optimized. The optimal linear transform will depend on the distortion measure selected. Further, the optimal transformation resulting from the selection of the distortion measure must produce decorrelated transform coefficients to validate the basis restricted transform coding approach. Minimization of the mean square error (MSE) distortion measure results in an optimal transform with this property. The selection of the MSE distortion measure is desirable because it maximizes the signal-to-noise ratio (SNR) for each block. This correlates well with the perceptual speech quality. The MSE is minimized on a block-by-block basis. Thus the transform coding scheme maximizes the segmental SNR where the segments are the analysis frames. Segmental SNR is a better indicator of the perceptual quality of speech than SNR. Nevertheless, selection of the MSE distortion measure may result in introduction of unacceptable perceptual distortions into the coded speech. Perceptual factors can be taken into account by modifying the bit assignment.



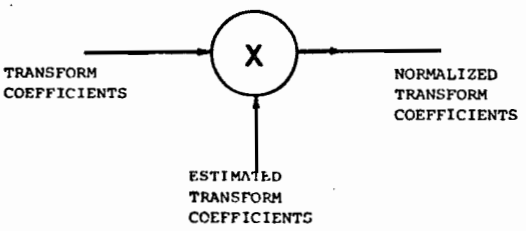
A) TRANSFORM COEFFICIENT DISTRIBUTION



B) PRE-SCALED TRANSFORM COEFFICIENT DISTRIBUTION
(good estimate)



C) PRE-SCALED TRANSFORM COEFFICIENT DISTRIBUTION
(better estimate)



PRE-SCALING EFFECT

FIGURE 2-2

2.3 OPTIMAL BIT ASSIGNMENT

The transform coefficient y_i with variance σ_i^2 requires coding with R_i bits/sample if the mean squared distortion D_i is not to be exceeded. R_i is given by:

$$\text{eq. 2.4} \quad R_i = \delta + \tfrac{1}{2}\text{LOG}_2 \left\{ \frac{\sigma_i^2}{D_i} \right\}$$

The second term is the minimal rate for independent identically distributed Gaussian random variables. The correction factor δ depends on the type of quantizer and the probability density function (pdf) of the signal. Neglecting the dependence of R_i on δ and substituting for D_i the optimal number of bits for quantizer Q_i is found by minimizing the average distortion given by

$$\overline{D} = \frac{1}{N} \sum_{i=1}^{N} D_i$$

with the constraint of a fixed average bit rate i.e.

$$\overline{R} = \frac{1}{N} \sum_{i=1}^{N} R_i$$

The expression for the average distortion is minimized subject to the constraint of a fixed bit rate by treating R_i as a continious variable and using an undetermined multiplier β one obtains:

$$\text{eq. 2.5} \quad \frac{\partial}{\partial R_i} \left\{ \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2 \, e^{-2R_i \ln 2} + \beta \sum_{i=1}^{N} R_i \right\}$$

$$\left\{ -\frac{1}{N} \sigma_i 2\ln 2 \; e^{-2R_i \ln 2} + \beta \right\} = \emptyset$$

It follows that

$$\text{eq. 2.6} \quad \sigma_i \, e^{-2R_i \ln 2} = \frac{N\beta}{2\ln 2} = \text{Const for } i=1,2\ldots N$$

Solving equation 2.6 for R_i and evaluating C using the constraint of a fixed bit rate we obtain:

$$\text{eq. 2.7} \quad R_i = \overline{R} \, \text{LOG}_2 \left\{ \frac{\sigma_i^2}{\left\{ \prod_{i=1}^{N} \sigma_i^2 \right\}^{\frac{1}{N}}} \right\}$$

Given an optimal bit assignment, the lower bound on distortion is given by

$$\text{eq. 2.8} \quad D_{lb} = 2^{2\delta} 2^{-2\overline{R}} \left\{ \prod_{i=1}^{N} \sigma_i^2 \right\}^{\frac{1}{N}}$$

Distortion introduced by a transform coding scheme depends on the distribution of the variances. In particular, the average distortion $\overline{D}$ is determined by the geometric mean of the variances.

The optimal quantization scheme discussed above was originally presented in a paper by Huang and Schultheiss [2]. Fractional and even negative bit assignments can result because R_i is treated as a continous variable. Segall [3] discusses the optimal bit assignment under the constraint of positive integer bit assignment.

Bits are assigned optimally in the Fortran simulation developed for this thesis. The technique uses k bit quantizers for the transform coefficients Y_k . The resulting mean square errors D(k) are tabulated in Max's paper [10]. It follows that the marginal return for the ith transform coefficient can be defined as

$$\text{eq. 2.9} \quad R_{i,k} = \sigma_i^2 \{D(k) - D(k+1)\}$$

Arranging $R_{i,k}$ in descending order, and assigning bits one-by-one, the global minimum mean square error will be achieved independently of the

distribution of the transform coefficients.

2.4 THE KARHUNEN-LOEVE TRANSFORM

Consider a discrete signal of N sampled values. This signal can be represented as point in an N dimensional space. Each sampled value is then a component of the N vector $\underline{X}$ which represents the signal in this space. Next consider a unitary transform (T) operating on the data vector $\underline{X}$ resulting in the transform vector $\underline{Y}$. The objective in data compression is to select a subset of M components of $\underline{Y}$ where M is less than N. The remaining components are discarded and this introduces some distortion. A unitary transform which minimizes the mean square error caused by discarding components is the objective. Some of the important properties of the KLT are described below.

The KLT is a data dependent transformation whose basis vectors are eigenvectors of the autocorrelation matrix of the X process. This transform diagonalizes the autocorrelation matrix of the transformed vector $\underline{Y}$ which means the components are uncorrelated and by the Gaussian assumption independent. Each transform coefficient can then be quantized independently without losing performance. It is possible to approximate $\underline{Y}$ by $\underline{Y}'$ in a lower dimensional space by discarding components. The mean square error resulting from this approximation is the sum of the variances of the discarded transform coefficients. If only the components of $\underline{Y}$ with the lowest variances are discarded, the approximation is optimal in a mean square error sense.

Two limitations of the Karhunen-Loeve transform are, that it is computationally burdensome and, requires solutions of eigenvector problems whose solutions may be numericaly unstable. More precisely the KLT requires a knowledge of the correlation function at the receiver to perform an inverse transformation. The correlation function is not generally available at the receiver.

2.5 SUB-OPTIMAL TRANSFORMS

The practical limitations of the KLT require an investigation of sub-optimal transforms. The discrete Fourier transform (DFT), the discrete cosine transform (DCT) and the Walsh-Hadamard transform (WHT) are all useful sub-optimal transforms. A method to compare the performances of unitary transforms in transform coding applications is highly desirable.

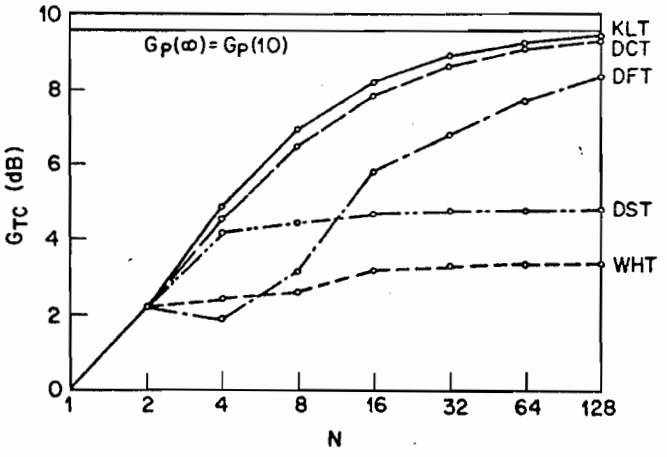
Assuming the transform only affects the probability density function (pdf) of the sampled process slightly, the quantizer parameter δ is unchanged whether quantizing in the time or transform domain. Hence the dependence of the distortion on δ is the same in either domain. From equation 2.4, it is clear that a lower bound on the distortion of a PCM is given by

$$\text{eq. 2.8} \quad D_{pcm} = 2^{2\delta} 2^{-2\bar{R}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2$$

Defining the transform coding gain over PCM as the increase in SNR over PCM, will enable useful comparisons.

$$\text{eq. 2.9} \quad G_{tc} \triangleq \frac{D_{pcm}}{D_{tc}} = \frac{\frac{1}{N} \sum_{i=1}^{N} \sigma_i^2}{\left\{ \prod_{i=1}^{N} \sigma_i^2 \right\}^{1/N}}$$

The transform gain for any unitary transform is then the ratio of the arithmetic and geometric mean of the variances of the transform coefficient as given above. Variances are just the diagonal elements of the co-variance matrix in the transform domain. Noll [4] modelled the long term statistics of voiced speech by a stationary tenth order Markov source. These results are shown in Figure 2-3. and illustrate transform coding gain dependence on block length N. For larger block lengths, the KLT performance improves. The DCT performs nearly as well. The DFT and DCT converge to KLT performance for a large block length. The discrete slant transform (DST) and WHT show a very poor performance in transform coding applications. Distribution of the variances of the discrete cosine transform coefficients are known to converge asymptotically to the power density spectrum of the process [4]. This property is used in the LPC adaptive bit assignment algorithm.



SUB-OPTIMAL TRANSFORM PERFORMANCE

FIGURE 2-3

2.6 THE DISCRETE COSINE TRANSFORM

For any practical transform coding implementation, the computational savings offered by the DCT and its near optimal performance, make it an attractive alternative to the KLT. The DCT and KLT are asymptotically equivalent if the data covariance matrices are Toeplitz hence it is not surprising that the DCT performs nearly as well as the KLT when the data vectors are large [11]. The rate distortion criteria of the KLT and DCT are also comparable [12].

Formally, the DCT of a real M point sequence v(n) can be defined as

$$V_c(k) = \sum_{n=0}^{M-1} v(n)c(k)\cos\{(2n+1)\pi k/2M\}$$

eq. 2.10

$$c(k)=\begin{cases} 1 & k=0 \\ \sqrt{2} & k=1,2,3,\ldots,M-1 \end{cases}$$

The inverse DCT is given by

eq. 2.11

$$v(n) = \frac{1}{M}\sum_{k=0}^{M-1} V_c(k)c(k)\cos\{(2n+1)\pi k/2M\}$$

The DCT of a sequence v(n) is closely related to a 2M point DFT of a related sequence u(n). Interpretation of perceptual information including formant structure and pitch striations in the Fourier transform domain may therefore be extended to the cosine transform domain. The following analysis follows Tribolet and Crochier[5,6]. Consider a 2M point sequence u(n) such that

eq. 2.12

$$u(n) = \begin{cases} v(n) & 0 < n < M-1 \\ 0 & M < n < 2M-1 \end{cases}$$

The 2M point DFT of u(n) is

$$U(k) = \sum_{n=0}^{M-1} u(n)e^{-j(2\pi kn/2M)}$$

eq. 2.13

$$= e^{j(k\pi/2M)} \sum_{n=0}^{M-1} u(n)e^{-j(\pi k(2n+1)/2M)}$$

$$k = 0,1,\ldots,2M-1$$

From the definition of the DCT it can seen that the DCT of v(n) denoted $V_c(k)$ is expressable in terms of U(k)

$$V_c(k) = \mathrm{Re}\{c(k)e^{-j(\pi k/2M)}U(k)\}$$

eq. 2.14

$$k = 0,1,2,\ldots,M-1$$

Denoting $|U(k)|$ and $\theta(k)$ as the magnitude and phase of U(k) equation 2.15 is obtained.

eq. 2.15 $V_c(k) = c(k)\,|U(k)|\,\cos\{\theta_k - \pi k/2M\}$

Clearly the DCT has an identical spectral envelope to that of the 2M point DFT. Thus, the DCT exhibits all properties of formant structure and pitch striations of the DFT.

The DCT reduces end-effect problems while maintaining minimal spectral redundancy [5,6]. This is another reason for its selection in transform coding. The superior performance of the DCT in this respect is a result of its close relationship to the 2M point DFT of a

sequence y(n). The sequence is formed from the M point sequence v(n) by defining

$$\text{eq 2.16} \qquad y(n) = \begin{cases} 1/2\ v(n) & n = 0,1,\ldots,M-1 \\ 1/2\ v(2M-1-n) & n = M,M+1,\ldots,2M-1 \end{cases}$$

The sequence y(n) is shown in Figure 2-4. The 2M point DFT of y(n) is given by

$$Y(k) = \sum_{n=0}^{2M-1} y(n) e^{-j(2\pi kn/2M)}$$

$$\text{eq. 2.17} \qquad = e^{j(\pi k/2M)} \sum_{n=0}^{M-1} v(n) \cos\{(2n+1)\pi k/2M\}$$

Comparing equations 2.17 and 2.10 the DCT of v(n) can be obtained from Y(k) according to the relation

$$\text{eq. 2.18} \qquad V_c(k) = c(k) e^{-j(\pi k/2M)} Y(k)$$

$$k = 0,1,2,\ldots,M-1$$

Although more computationally efficient algorithms have been described in the literature [13,14,15], the above interpretation proves useful in understanding the end-effects which in transform coding can cause undesirable clicking at the block rate. In low bit rate coding the quantization noise is a combination of both multiplicative and additive effects.

$$\text{eq. 2.19} \qquad \hat{V}(k) = G(k)\ V(k) + E(k)$$

where $\hat{V}(k)$ are the quantized Fourier transform coefficients and $G(k)$ and $E(k)$ are the noise terms. For low bit rate coding some values of the DCT may not be encoded leading to a low-pass effect. The synthesis procedure leads to the result

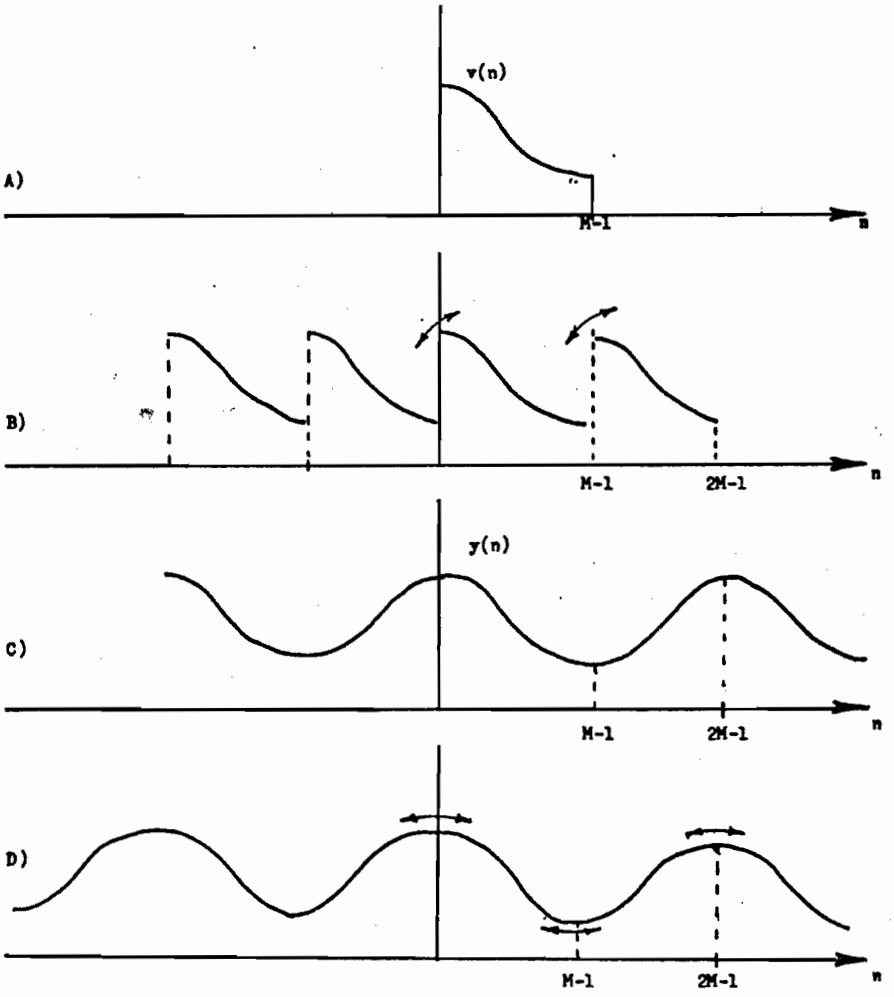
eq. 2.20 $$\hat{v}(n) = v(n) \otimes g(n) + e(n)$$

where $g(n)$ and $e(n)$ are the inverse transforms of G and E respectively. The circular convolution denoted by $\otimes$ results in aliasing in time as illustrated by the arrows in Figure 2-4b and, is the cause of the edge effect artifacts. Because of association of the DCT with the synthesis sequence $y(n)$, quantizing $V_c(k)$ is equivalent to quantizing $Y(k)$, therefore

eq. 2.21 $$\hat{Y}(k) = G(k)\ Y(k) + E(k)$$

$$\hat{y}(n) = g(n) \otimes y(n) + e(n)$$

Circular convolution still causes aliasing in time but the effects are not so severe. Thus the DCT produces less noticeable boundary effects than the DFT in transform coding applications.



A ONE DATA BLOCK v(n)

B END EFFECTS OF CIRCULAR CONVOLUTION FOR DFT

C SYMMETRIC SEQUENCE y(n)

D END EFFECTS OF CIRCULAR CONVOLUTION FOR DCT

BLOCK BOUNDARY DISTORTION

FIGURE 2-4

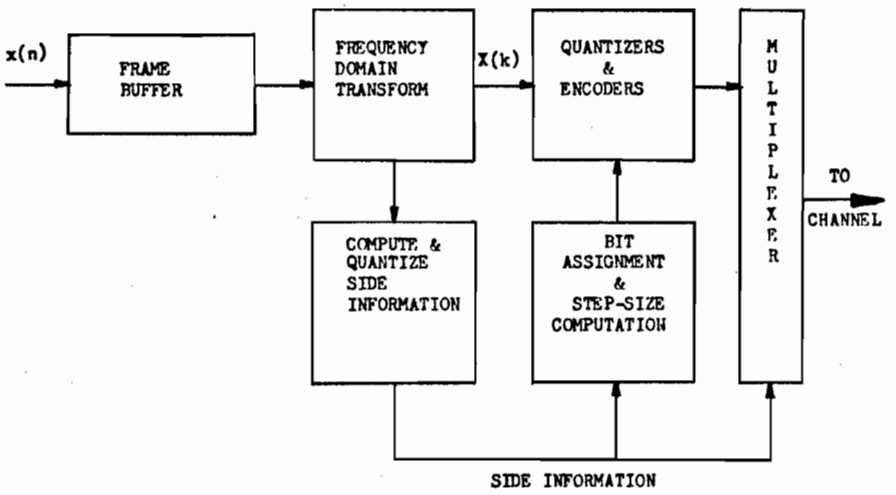
CHAPTER 3

ADAPTIVE TRANSFORM CODING

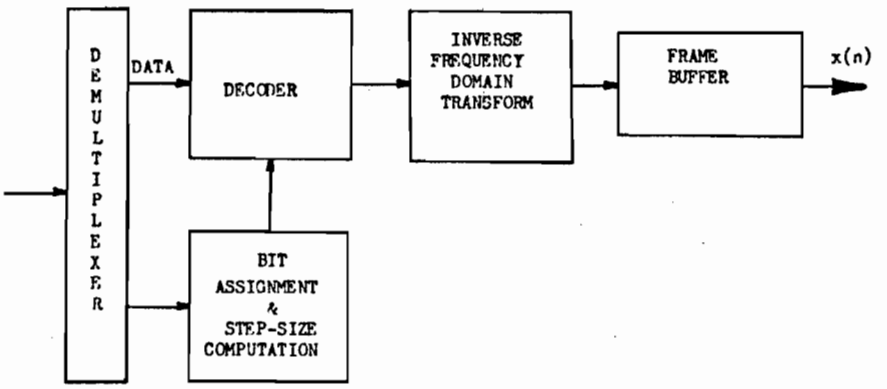
The general structure of adaptive transform coding is shown in Figure 3-1. Speech is sampled, quantized, buffered into analysis frames, and transformed. Transform coefficients are adaptively quantized by uniform quantizers and transmitted to the receiver. The quantizers are characterized by the number of quantization levels and their step size. The step size is determined from an estimate of the spectral variance of the transform coefficients. The number of quantization levels to be used for each coefficient is determined from the speech spectrum for each frame through a bit assignment process. The distribution of the transform coefficient variances, which control step size adaptation, and bit assignment, is known as the basis spectrum. The basis spectrum is parameterized, encoded, quantized and transmitted to the receiver as side information. For a 9.6 kb/sec coder, 2 kb/sec is considered adequate to transmit all the side information to the receiver. The receiver then has knowledge of the bit assignment and can reconstruct the transform coefficients. Inverse transformation recovers the speech signal. Bit assignment and step size adaptation are collectively known as the quantization strategy and determine the distribution of quantization noise in the

frequency domain. The coder design process procedes in two steps. The first step is to maximize the SNR. The second step modifies the quantization strategy according to perceptual criteria. In particular, auditory masking principles can be applied to obtain better subjective performance from the transform coder.

Three transform coding schemes are discussed in the following sections. These use different spectral parameterization techniques. Their structure and performance are analyzed in the context of what has been discussed previously.



A) TRANSMITTER



B) RECEIVER

GENERAL STRUCTURE OF ADAPTIVE TRANSFORM CODING

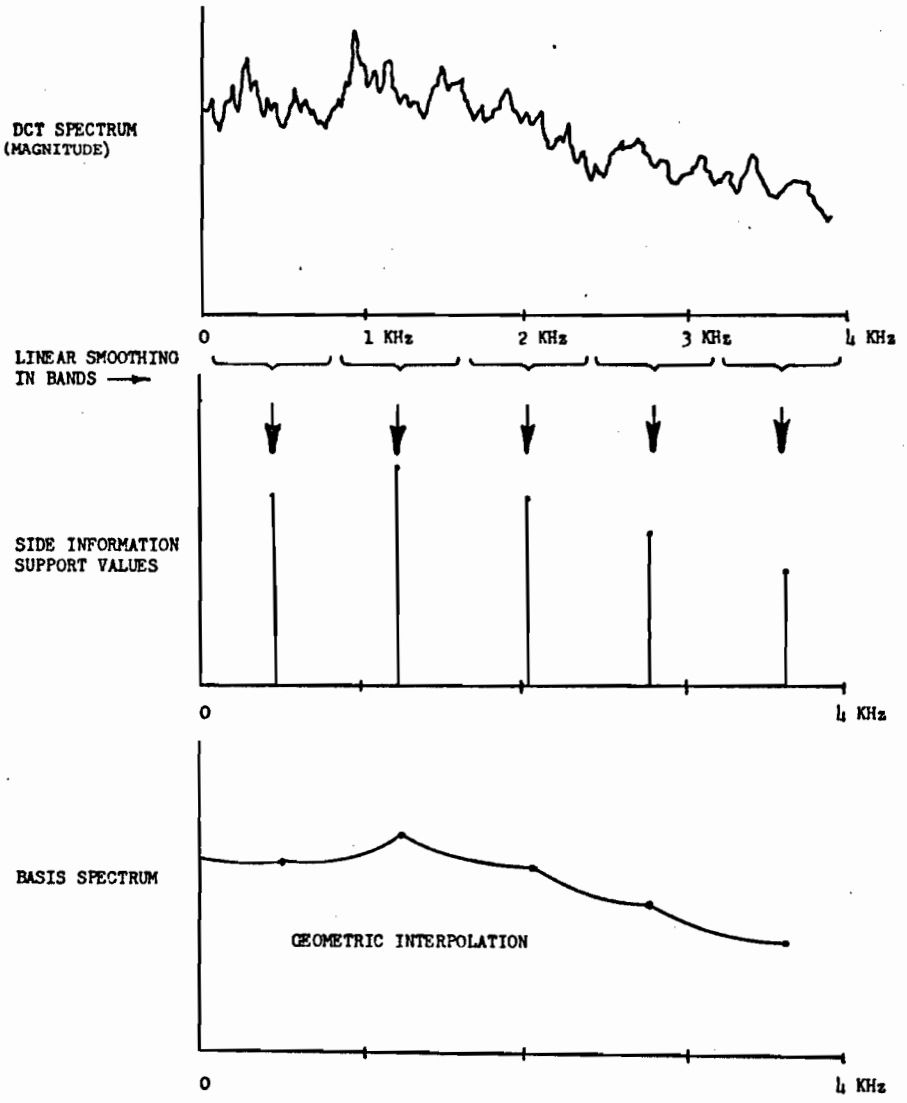
FIGURE 3-1

3.1 LOG-LINEAR SMOOTHING TECHNIQUE

The log-linear smoothing technique is a method of transmitting the necessary side information for step size allocation and bit assignment by spectral parameterization [4]. An estimate of the transform coefficient variance is obtained by squaring the DCT coefficients for each analysis frame. Correlation between adjacent transform coefficients is exploited by averaging. This reduces the number of squared values to be transmitted. The remaining values are called basis spectrum support values and constitute side information. The logarithms of the support values are linearly interpolated at the receiver. In the linear domain this is equivalent to geometric interpolation and gives the basis spectrum. This is an estimate of the spectral variance of the transform coefficients. When using the optimal transform, (the KLT,) the transform coefficients should have no correlation. However, the side information parameterization scheme exploits any correlation present to obtain a compressed estimate of the variance of the transform coefficients.

The process of spectral parameterization is shown in Figure 3-2. The smoothing inherent in this technique annihilates the fine line structure of the spectral variance. Pitch information is lost as a result. Improved fit to the spectral variance in the mean square error sense can be achieved if more sophisticated methods are used. For these reasons LPC adaptive transform coders have recently received more study than those using the log-linear smoothing technique. Zelinski and Noll [4] proposed many refinements to the smoothing technique in addition to new approaches for quantizing the side information. Use of LPC smoothing was also suggested. Sloan [16]

made detailed experimental comparisons of the two methods. His results indicate that the all-pole model of the short term spectrum results in better speech reproduction. Refinements in the coder such as pitch modelling and quantization noise shaping are accomplished more easily using the all-pole model.



LOG – LINEAR SMOOTHING

FIGURE 3–2

3.2 ALL-POLE MODEL

The log-linear smoothing technique is a non-speech specific method for parameterizing the basis spectrum. A full knowledge of models and dynamics of speech production can be exploited by using an algorithm based on an all-pole model of the formant structure of speech. In conjunction with a pitch model representing the fine structure, the algorithm can be particularly effective in transform coding. The resulting system is referred to as a LPC adaptive transform coder.

Figure 3-3 illustrates the block diagram of the coder under discussion. Speech is buffered into analysis frames typically ranging in length from 32 to 256 samples, for 8 khz speech 4 to 24 ms. The buffered speech is then transformed. The DCT is used because it is a good overall fixed transform. Spectral coefficients are squared to obtain a measure of the spectral variance of the DCT coefficients. The basis spectrum is obtained through a LPC analysis as follows. An autocorrelation-like function, referred to as the pseudo-autocorrelation function (pseudo-ACF), is derived by an inverse Fourier transforming the squared DCT coefficients. The pseudo-ACF exhibits properties similar to the normal ACF because of the DCT relationship to the Fourier spectrum. Values of the autocorrelation function are used to solve the LPC autocorrelation equations. The solution is expressed for convience in the form of a gain factor and a set of reflection coefficients. These are quantized and transmitted to the receiver. The inverse LPC spectrum is an estimate of the basis spectrum. The fine structure of the DCT spectrum is obtained from the pitch model. The pitch period is derived by searching the pseudo-ACF

for the first maximum greater than the order of the predictor. The corresponding pitch gain is the ratio of the values of the ACF at the pitch period over its value at the origin. With these two values, a pitch pattern can be generated in the frequency domain. The model is of the form [6]

$$\text{eq. 3.1} \quad \sigma_p = \left\| \frac{1}{1-Ge^{-j\omega L}} \right\| * H[\omega]$$

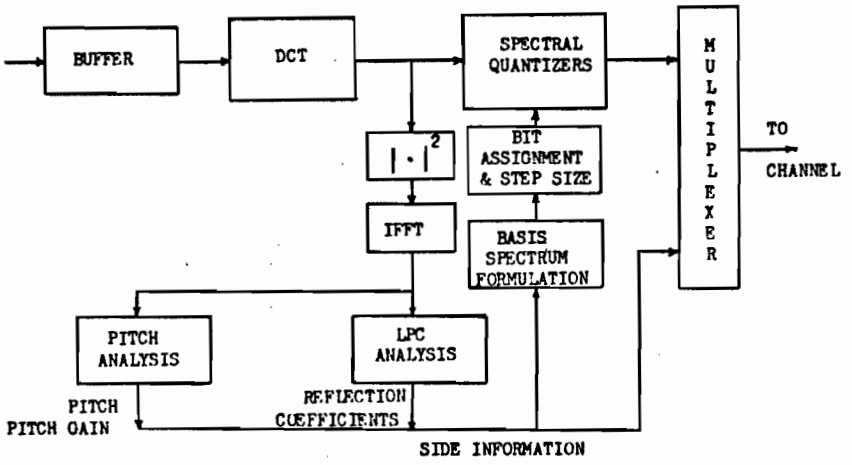
and is associated with a windowed, one sided periodic impulse train with exponential decaying amplitudes. In the time domain this corresponds to

$$\text{eq. 3.2} \quad p[n] = h[n] \cdot \sum_{m=0}^{\infty} G^m \, \delta[n-mL]$$

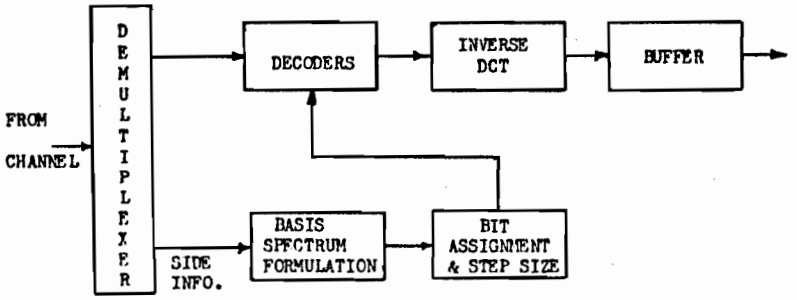
L pitch period

G pitch gain

where h(n) is the analysis window. The Fourier transform of the pitch time sequence is calculated. Frequency representation of the pitch is impressed on the LPC estimate to obtain the basis spectrum used in step size adaptation and bit assignment. This procedure is illustrated in Figure 3-4. Pitch gain and predictor error are quantized and transmitted to the receiver using a mu-law quantizer.



A) TRANSMITTER

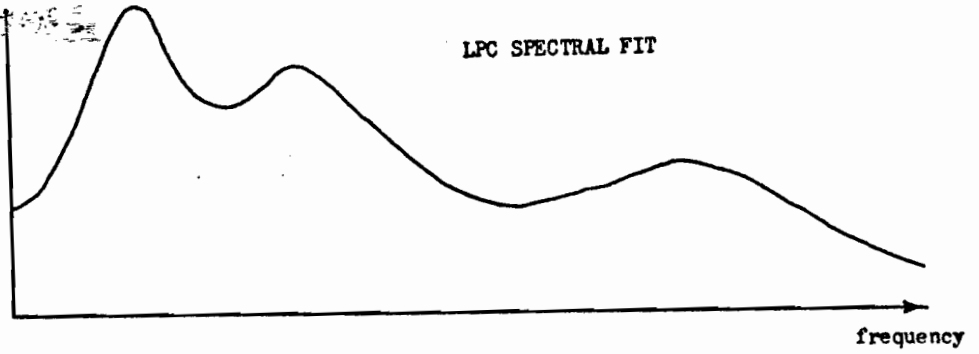


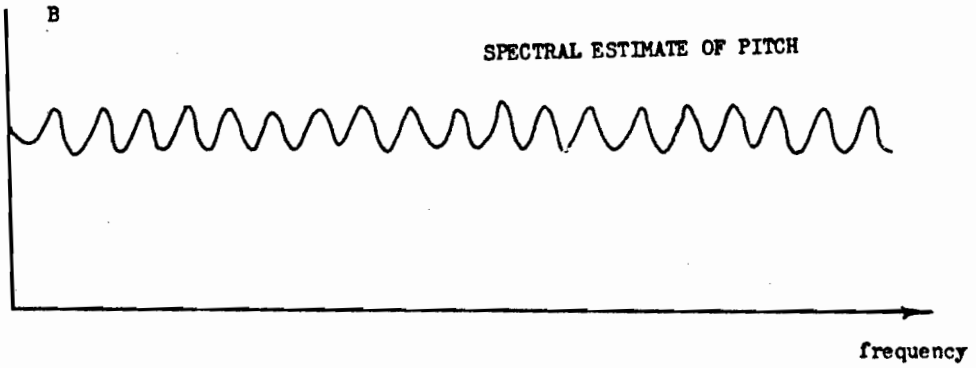
B) RECEIVER

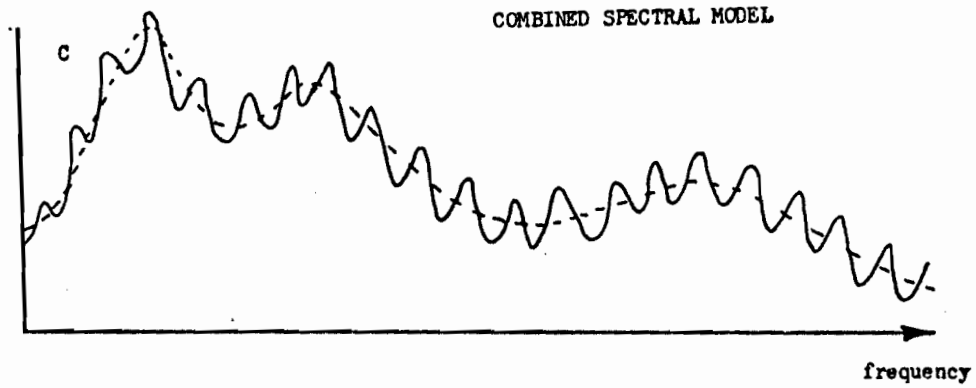
LPC ADAPTIVE TRANSFORM

CODER

FIGURE 3-3







LPC PITCH MODEL

FIGURE 3-4

3.3 HOMOMORPHIC MODEL

The homomorphic model under discussion here is also a model appropriate to speech with properties similar to that of the LPC and pitch models. Figure 3-5 shows the most important aspects of the homomorphic method of parameterizing the speech spectrum. The log-magnitude of the DCT coefficients are computed and inversely transformed. The resulting sequence c(n) is known as the cepstrum. In conventional homomorphic analysis, the transform coefficents are obtained by a Hamming window DFT analysis. In transform coding, the speech spectrum is acquired by a DCT analysis using rectangular or raised cosine windowing. This does not result in as good a spectral estimate as the DFT approach. It is for this reason that c(n) will be referred to as the pseudo-cepstrum. Formant and pitch structure of speech can be extracted from the pseudo-cepstrum. Formant parameters are extracted by cepstral windowing and pitch parameters by peak picking. Formant and pitch parameters are encoded and transmitted to the receiver as side information. The cepstral model is reconstructed by combining the formant and pitch parameters to obtain a cepstral representation c(n). The desired log-spectrum is obtained through inverse transformation.

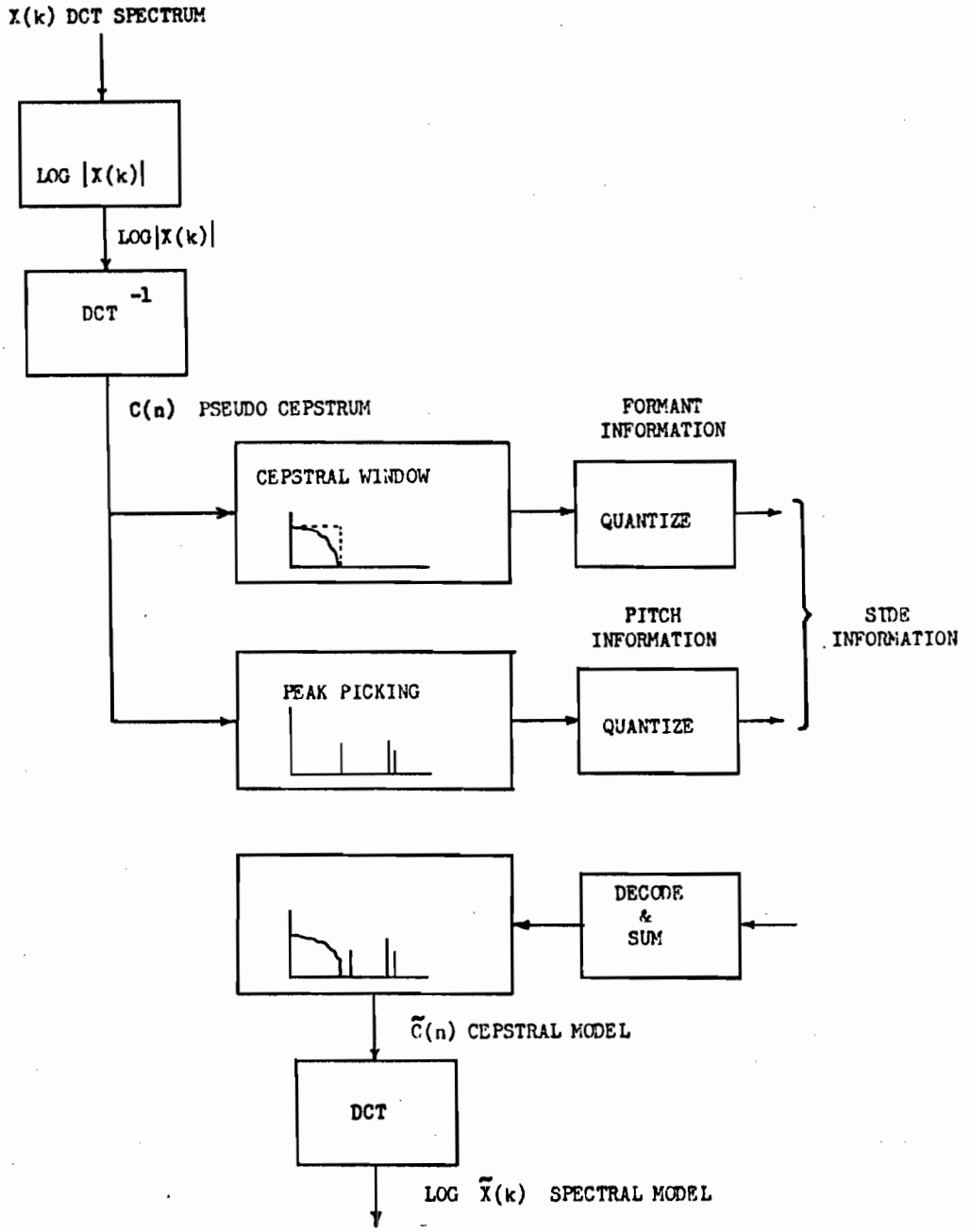
Formant structure of speech manifests itself in the first 10 to 25 coefficients of the cepstrum. By windowing the low-time region of the cepstrum and quantizing these parameters, a cepstral representation of the envelope of the transform coefficients can be obtained. The log-magnitude of a speech spectrum can be obtained by decoding the windowed cepstral coefficients, padding the balance with zeros and transforming with a DCT. The purpose of both the LPC and

homomorphic models is to extract the formant envelope. These models give roughly equivalent mean square error performance in estimating the actual smooth component of the spectrum [8]. The LPC spectral estimate does a better job in the formant regions, but is less accurate in the low amplitude regions. The homomorphic model distributes bits more evenly in the frequency domain than the LPC model. Considerations in hardware implementation of an adaptive transform coder may favour one technique over another.

Pitch structure of speech exhibits itself as a periodic train in the cepstrum. Harmonic pitch structure in the spectrum can therefore be obtained by measuring the location and amplitude of the pulses in the cepstrum by peak picking. Since amplitude of these pulses decays rapidly with increasing time, it is sufficient to measure the location and amplitude of only the first one or two pulses. This information is then quantized and transmitted as side information. The pitch model consists of two cepstral peaks. The primary pitch peak is located by searching the 2 ms to 16 ms (60-500 Hz) region of the cepstrum. The largest adjacent value to the primary pitch peak is located. This recognizes that the main cepstral peak may be more than one sample in width. A search is made for a secondary pitch peak in the region of half and twice the pitch value previously determined. Location and amplitude of these peaks are encoded and transmitted. An artificial gain factor of 1.5 is applied to all pitch amplitudes to help provide a better spectral match to the peaks of the actual spectra as suggested by Cox and Crochiere [8]. The reason for this is that homorphic analysis averages in the log domain which tends to lower pitch peaks. Introduction of this gain raises the entire spectral estimate such that the estimate matches the peaks in the

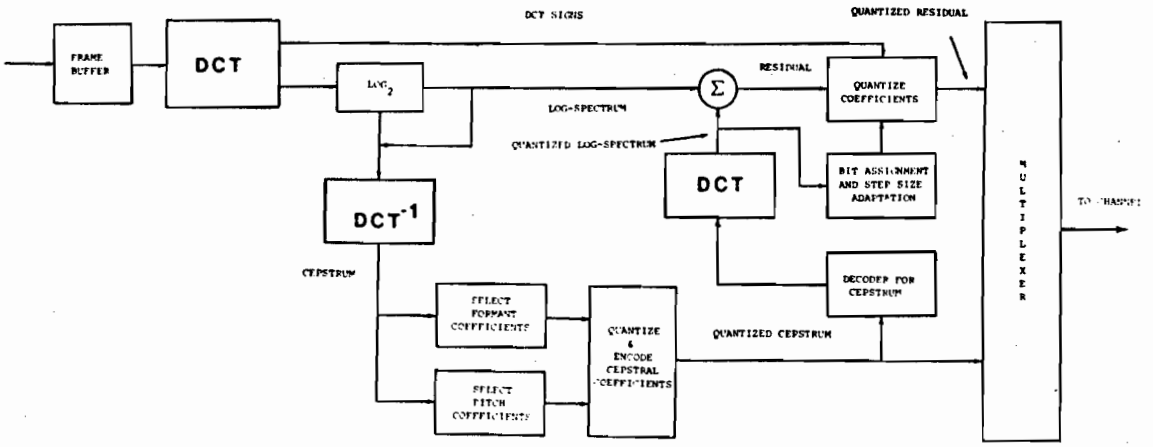
log-spectrum rather than peaks in the average log-spectrum.[8]

Figure 3-6 illustrates the block diagram of the coder under discussion. Speech is buffered into analysis frames and transformed. Spectral coefficients are used as a measure of the spectral variance of the DCT. The basis spectrum is obtained from a homomorphic analysis as follows. Formant and pitch cepstrum coefficients are isolated from the pseudo-cepstrum and result in a cepstral representation of the basis spectrum. Cepstrum coefficients are then transmitted as side information to the receiver. The locally decoded cepstrum is frequency transformed with a DCT to obtain the basis spectrum used for step size adaptation and bit assignment. The basis spectrum for the homomorphic model closely follows the basis spectrum for the LPC model. The difference of the log-spectrum and basis spectrum is transmitted to the receiver. Additional side information, namely the sign of the transform coefficients, is required for operation of this coder. The residual is decoded at the receiver by adding the basis spectrum to the residual and converting to the linear domain. The sign of the coefficients is introduced and inverse transformation recovers the speech waveform.

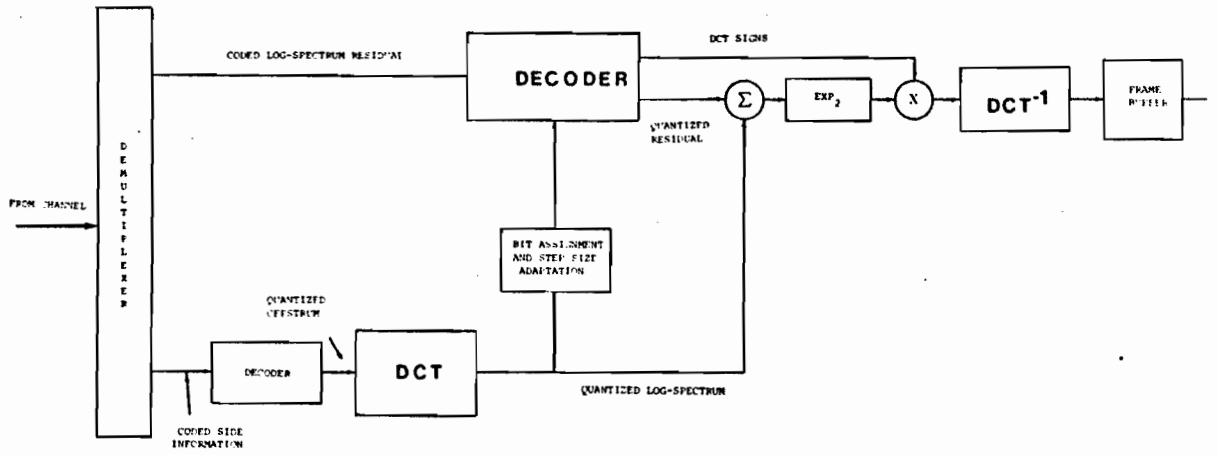


HOMOMORPHIC SIDE INFORMATION PROCESSING

FIGURE 3-5



A) TRANSMITTER



B) RECEIVER

HOMOMORPHIC ADAPTIVE TRANSFORM CODER STRUCTURE

FIGURE 3-6

CHAPTER 4

CODER EVALUATION

4.1 SIMULATION PROCEDURE

Homomorphic and LPC coding schemes have been simulated in software. Fortran simulation has been used to permit easy modification and enhancement of coder structure, and facilate experimental optimization of coder parameters. Trade-offs between coder complexity and performance can be investigated in conjunction with studies of the effect of quantizing and interpolating side information parameters. The complexity of the coder can be decreased by using shorter frame lengths. The performance of the coder should decline gradually according to Figure 2-3. However, this does not consider the effect of frame boundary discontinuities which is significent. Perceptual quality of transform coded speech can be improved by introducing pre-emphasis, spectral shaping, and pitch modelling. The perceptual quality can only be judged by actual listening tests. In summary, software simulation permits evaluation of subjective performance of adaptive transform coders.

The simulation procedure requires three steps. The first is to digitize a segment of speech and store the digitized signal (on disk). Unless otherwise stated, all source audio files are digitized at 8 Khz after passing through an anti-aliasing filter. The full 0-4 Khz bandwidth is coded. The second step simulates the coder by accessing the stored speech and producing coded speech which is also stored on disk. Coding is not done in real time. The third step accesses stored coded speech and regenerates this as an analog waveform in real time. The perceptual quality of the speech can now be judged. Simulation is performed on a Vax 11/780 computer. A large body of utility programs and a 15 bit A/D D/A converter combination permit real time speech acquition and play back.

Source coding and channel coding problems can always be isolated and solved separately. For this reason, the effect of channel error is not considered in this thesis. Goldberg and Cosell [9] have investigated channel error in the context of transform coding and has developed methods to combat this.

4.2 THE LPC CODER

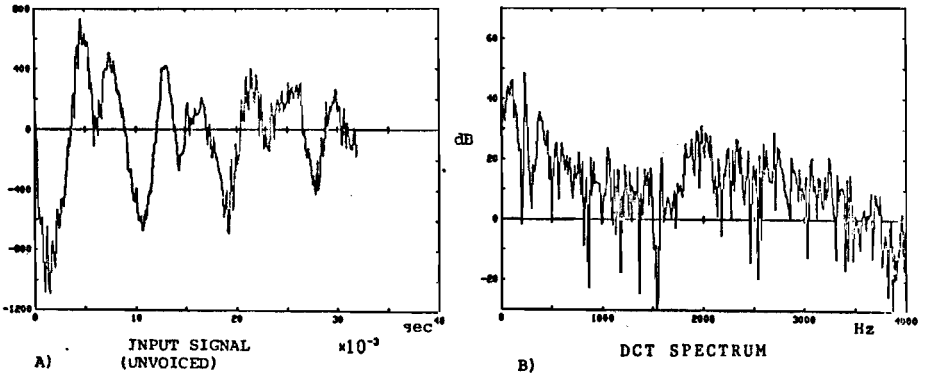
In this section the operation of the LPC Adaptive Transform Coder is discussed in more detail. Impairment in speech quality caused by the processing are enumerated and explanations sought. Subjective effects of pre-emphasis, spectral shaping and pitch modelling are presented. Degradation in coder performance due to parameter quantization is considered. Analysis frame windowing effects and their ability to combat frame boundary discontinuities are studied. Computational saving resulting from shorter frame lengths and

associated problems with side information interpolation and frame boundary discontinuities are considered.

4.2.1 Coder Operation

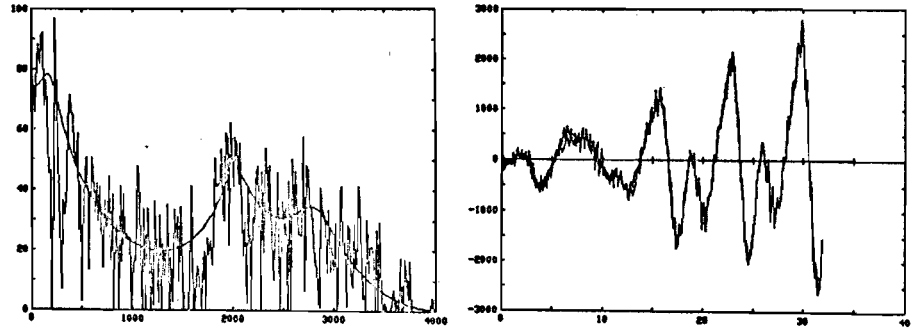
The structure of the LPC Adaptive Transform Coder is reviewed in section 3.2. Each waveform in the coding process of Figure 3-3 is illustrated in Figure 4-1. A typical 256 point unvoiced speech frame is shown in Figure 4-1a. A first step in the coding process is the frequency domain transform. The DCT of the speech frame and the squared DCT spectrum, (which is considered to be an estimate of the spectral variance of the DCT coefficients) are shown in Figures 4-1b and 4-1c respectively. The spectral estimate of the variance of the transform coefficients is parameterized by a LPC analysis. The LPC fit to the squared DCT spectrum is obtained by inverse Fourier transformation of the squared DCT spectrum to obtain the pseudo-autocorrelation function. These values are used to define the LPC auto-correlation normal equation. The inverse LPC spectrum is a fit to the spectral variance of the transform coefficients and, is compared to the squared DCT spectrum in Figure 4-1c. The basis spectrum is the quantized parameterized inverse LPC spectrum with impressed pitch information on which step size adaptation and bit assignment for the coefficient quantizers are based. Figure 4-1d shows a voiced input speech frame laid over the corresponding output speech frame. Figure 4-1e shows the basis spectrum for this analysis frame with and without pitch information incorporated. Figure 4-1f illustrates bit assignment for the unvoiced speech frame. The bit assignment is zero between 750-1750 Hz and 3000-4000 Hz. Coded speech for these frequency regions contains no energy. These dead frequency

bands will vary in width between adjacent frames causing a perceptable warble in the speech. Zelinski and Noll [4] recommend filling the dead regions with zero mean white noise with a variance weighted by the basis spectrum to combat the low-pass effect. Figure 4-1g shows the quantized DCT and Figure 4-1h gives a comparison of the original DCT and the quantized DCT. Figure 4-1i is the quantized DCT with the dead regions filled with noise to combat the low-pass effect. Figure 4-1j shows the original unvoiced speech frame and coded speech frame overlaid.



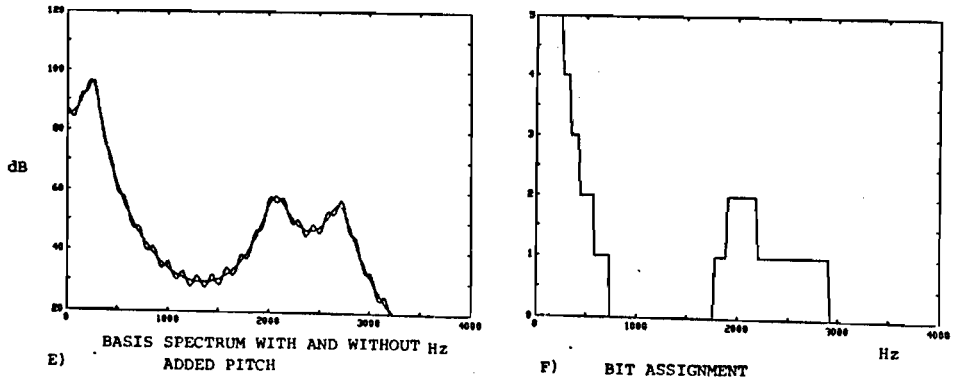
A) INPUT SIGNAL (UNVOICED) $\times 10^{-3}$

B) DCT SPECTRUM



C) SQUARED DCT SPECTRUM WITH BASIS SPECTRUM FIT

D) VOICED INPUT SIGNAL COMPARED TO OUTPUT SIGNAL



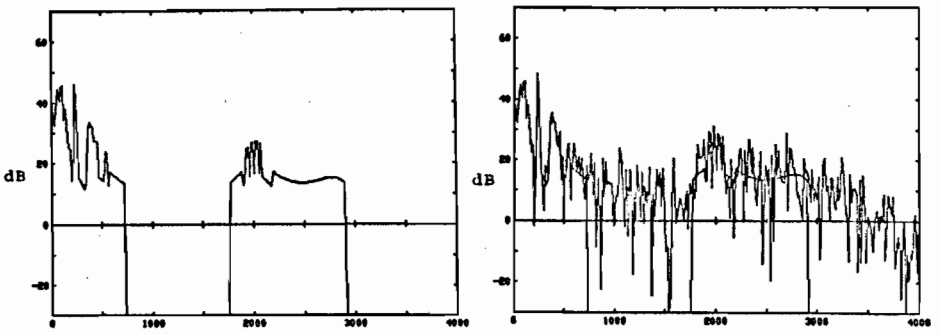
E) BASIS SPECTRUM WITH AND WITHOUT ADDED PITCH

F) BIT ASSIGNMENT

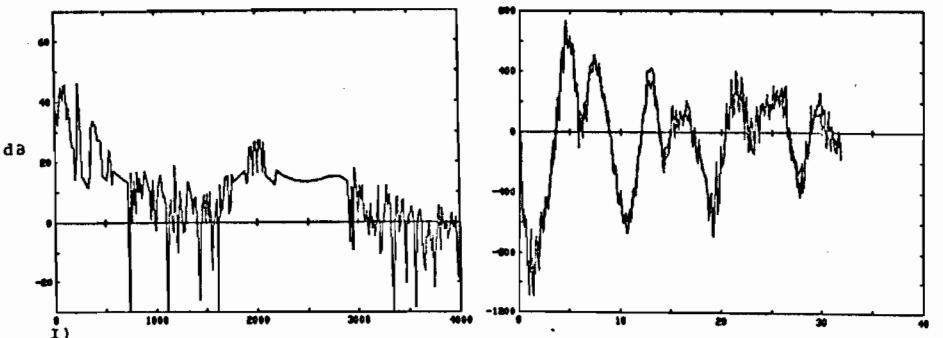
LPC ADAPTIVE TRANSFORM CODER

WAVEFORMS

FIGURE 4-1



G) QUANTIZED DCT H) QUANTIZED AND UNQUANTIZED DCT SPECTRA COMPARED



I) QUANTIZED DCT WITH NOISE FILLING J) INPUT AND OUTPUT SIGNALS COMPARED

LPC ADAPTIVE TRANSFORM CODER

WAVEFORMS

FIGURE 4-1

4.2.2 Reducing Transform Complexity

Adaptive Transform Coders offer high quality speech for coding rates between 10 Kb/sec and 20 Kb/sec. Reconstructed speech from these coders approaches the telephone toll standard (56 Kb/sec companded PCM). For higher rates, quality improves only gradualy. For lower rates, speech quality rapidly deteriorates. Adaptive transform coding is superior to most other speech coders in terms of speech quality for this bit rate range. Unfortunately, Adaptive Transform Coders are among the most complex of the speech coders. A reduction of complexity of transform coders while maintaining speech quality is an important objective of this research. Reduced complexity can be achieved by simplifing the side information spectral parameterization scheme or by simplifing the transform. The DCT is a good transform for coding applications because it combats frame boundary discontinuities and, has near KLT performance while being computationally simpler. Shortening frame length is the simplest technique to reduce transform complexity but has several consenquences. The theroretical performance of the DCT declines and frame boundary discontinuities become more frequent. Also, side information cannot be updated each frame and still be transmitted within the 2 kb/sec allocated. A side information interpolation scheme must be developed which compresses the transmission of reflection coefficients, pitch and gain parameters into 2 kb/sec. Side information interpolation results in an additional degradation of coder performance because the precision and, or frequency of parameter update must be reduced resulting in a poorer basis spectrum estimate of the transform coefficient distribution.

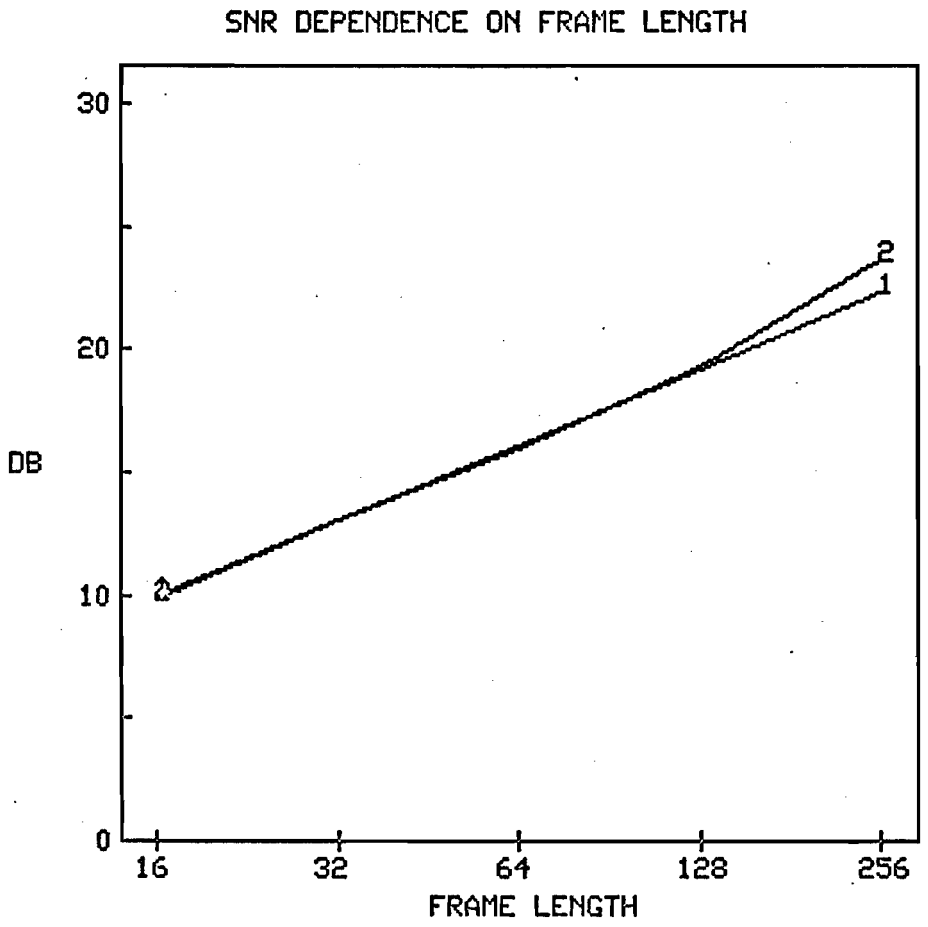
Figure 4-2 shows the SNR and segmental SNR performance of the LPC adaptive transform coder for various frame lengths. The input signal is a three second 1 Khz sinusoid and is coded with unquantized side information parameters. This signal is "stationary" and side information interpolation has no effect on performance of the coder. Frame boundary discontinuities are mainly a perceptual effect and are not responsible for the declining SNR performance. The predominate effect is a reduced performance of the transform described in section 2.4 and Figure 2-3.

Speech is a non-stationary signal and shorter frames may model the speech better despite the degradation caused by side informtion interpolation. Figure 4-3a to 4-3.9 shows the SNR and segmental SNR performance verses frame length with and without side information interpolation for several sentences. The side information is unquantized, and interpolated using the Zelinski and Noll [4] method discussed in Section 4.2.3. The phrases of a male speaker are listed in Table 1. Clearly the decreased performance of the transform is compensated by the better non-stationary speech modelling offered by shorter frame lengths.

TABLE 1

AUDIO SOURCE FILE	PHRASE
DOUG1	The birch cannoe slid on the smooth planks.
DOUG2	Glue the sheet to the dark blue background.
DOUG3	It's easy to tell the depth of a well.
DOUG4	These days a chicken leg is a rare dish.
DOUG5	Rice is often served in round bowls.
DOUG6	The juice of lemons makes fine punch.
DOUG7	The box of lemons was thrown beside the parked truck.
DOUG8	The hogs were fed chopped corn and garbage.

From these results it can be observed that side information interpolation causes up to a 4 dB drop in SNR performance. Side information interpolation is necessary in order to transmit side information within the allocated 2 Khz. Figure 4-4a to 4-4h show the SNR and segmental SNR performance for the adaptive transform coder when side information parameters are quantized to 64 levels each within the 2 Khz bandwidth. The SNR drops very quickly with decreased frame size. Statistics of the side information parameters are known to change with frame length. This is thought to be responsible for the declining SNR values shown in Figure 4-4. The consequence of this is that side information quantizers must be specifically designed for a specific parameter and frame length. The statistics of side information parameters will be investigated in a following section.

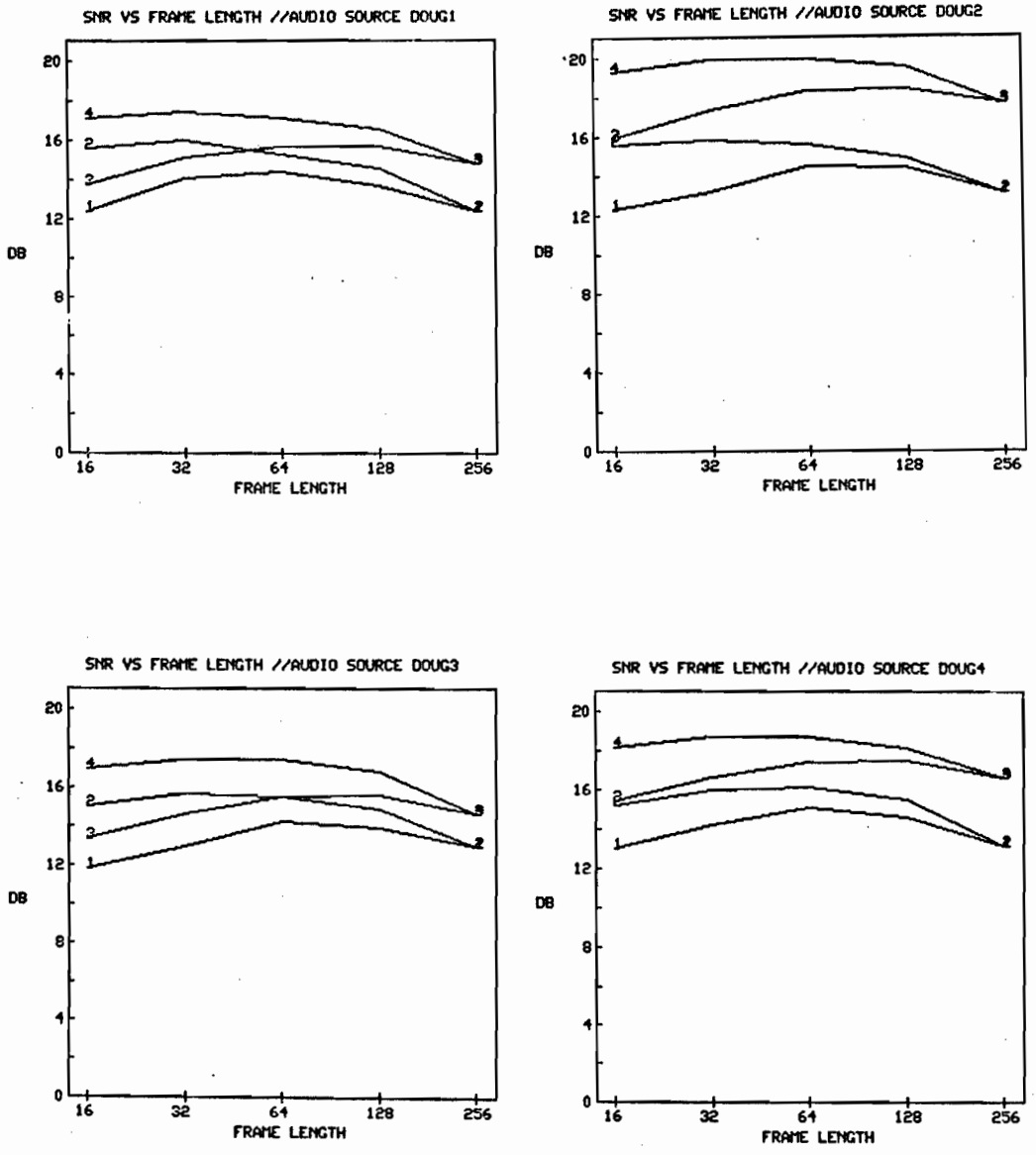


9.6 KB/SEC ATC CODER

1 SNR 1 KHZ TONE
2 SEG. SNR 1 KHZ TONE

CODER SNR PERFORMANCE

FIGURE 4-2

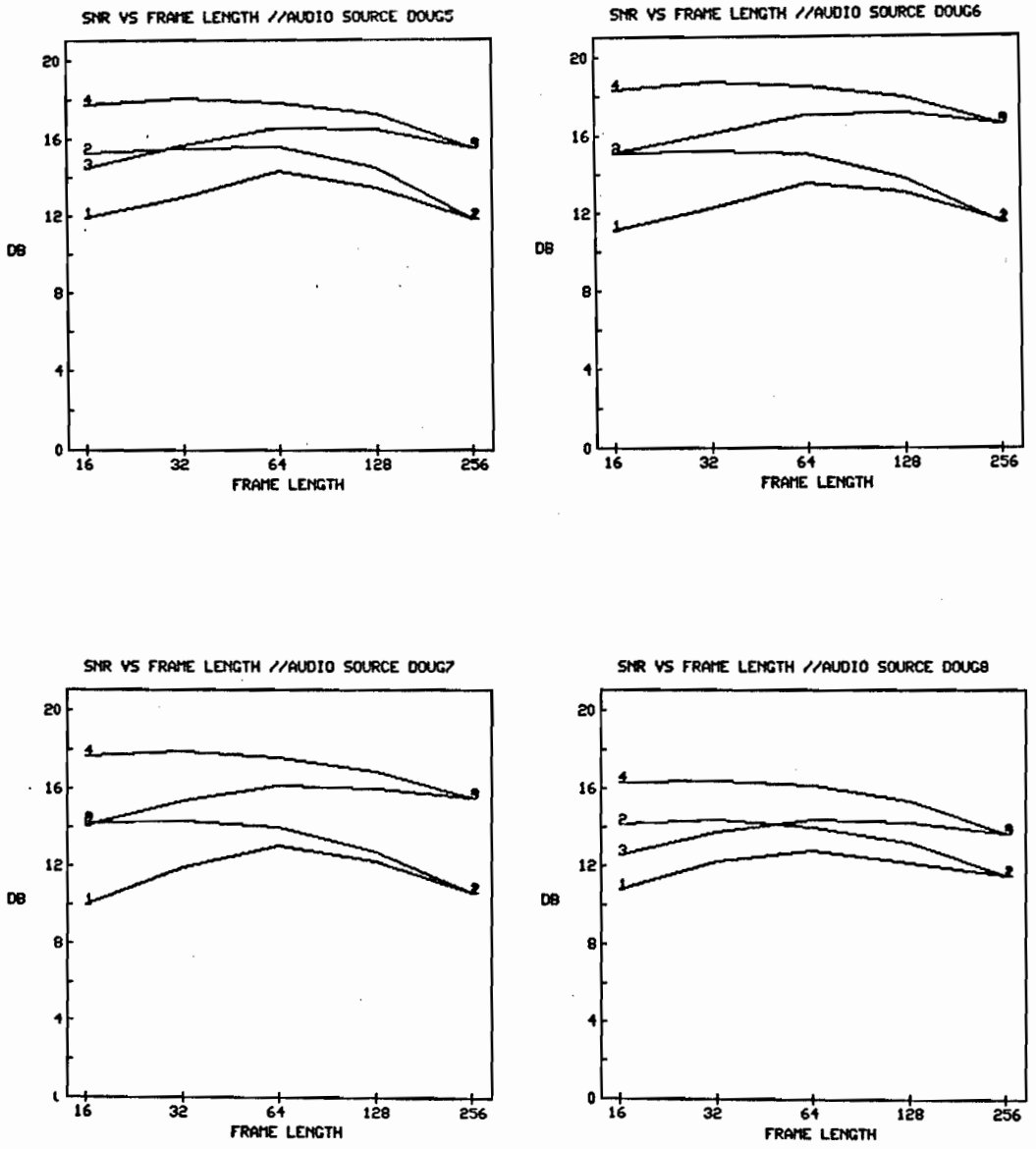


UNQUANTIZED SIDE INFORMATION
9.6 KB/SEC ATC CODER

1 SNR INTERPOLATING
2 SNR NONINTERPOLATING
3 SEG. SNR INTERPOLATING
4 SEG. SNR NONINTERPOLATING

CODER SNR PERFORMANCE

FIGURE 4-3

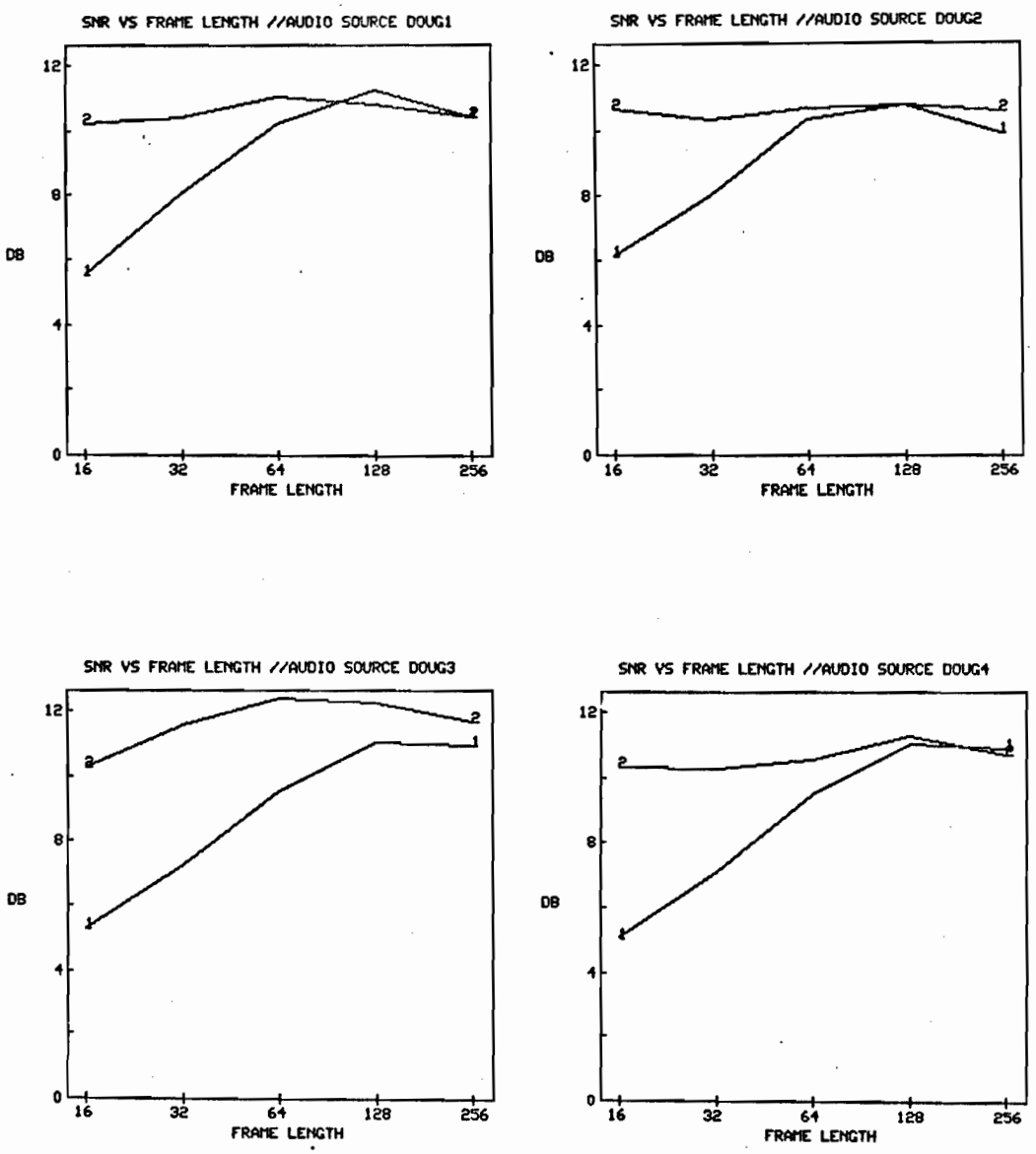


UNQUANTIZED SIDE INFORMATION
9.6 KB/SEC ATC CODER

1 SNR INTERPOLATING
2 SNR NONINTERPOLATING
3 SEG. SNR INTERPOLATING
4 SEG. SNR NONINTERPOLATING

CODER SNR PERFORMANCE

FIGURE 4-3

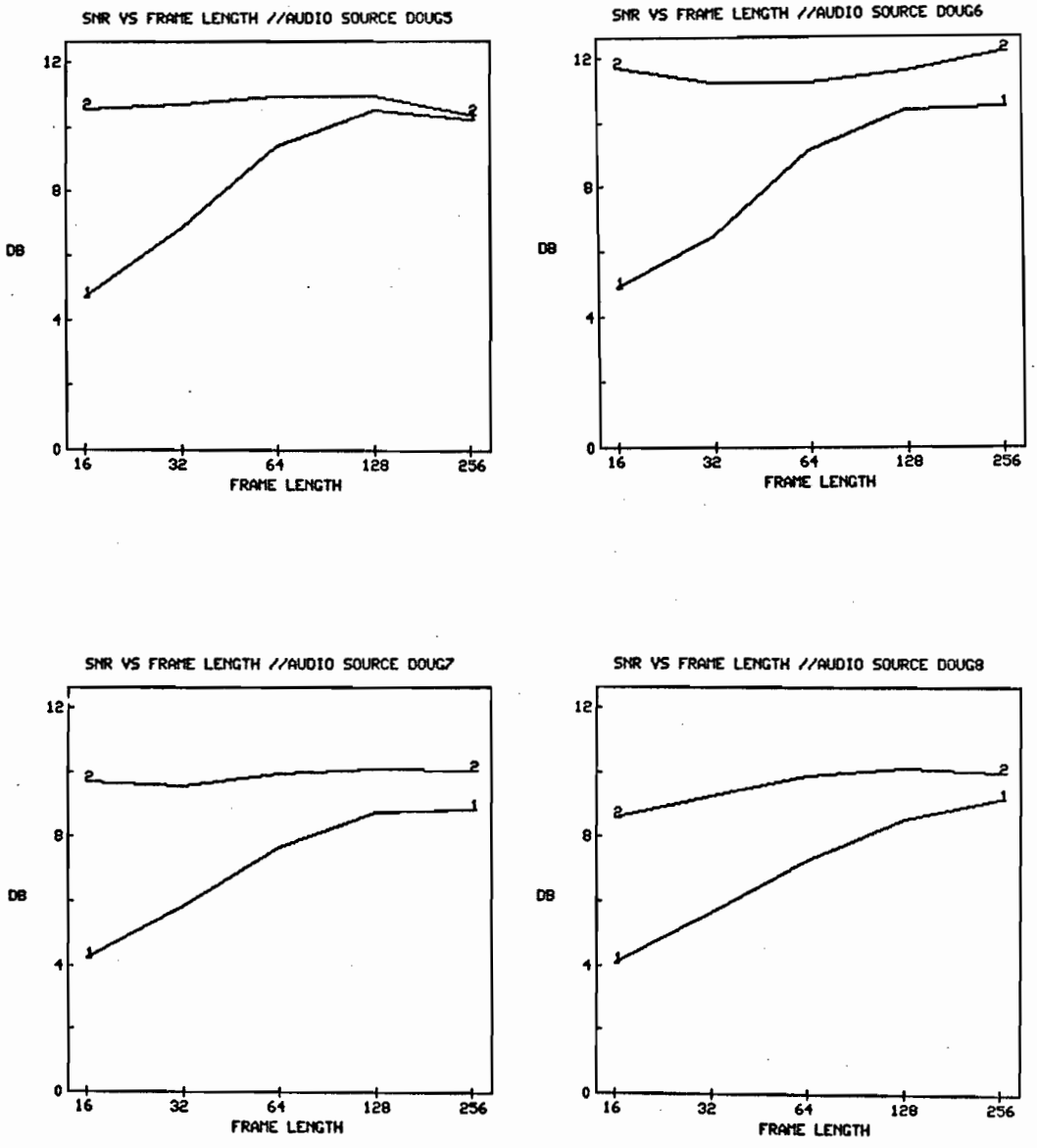


QUANTIZED SIDE INFORMATION
9.6 KB/SEC ATC CODER

1 SNR INTERPOLATING
2 SEG. SNR INTERPOLATING

CODER SNR PERFORMANCE

FIGURE 4-4



QUANTIZED SIDE INFORMATION
9.6 KB/SEC ATC CODER

1 SNR INTERPOLATING
2 SEG. SNR INTERPOLATING

CODER SNR PERFORMANCE

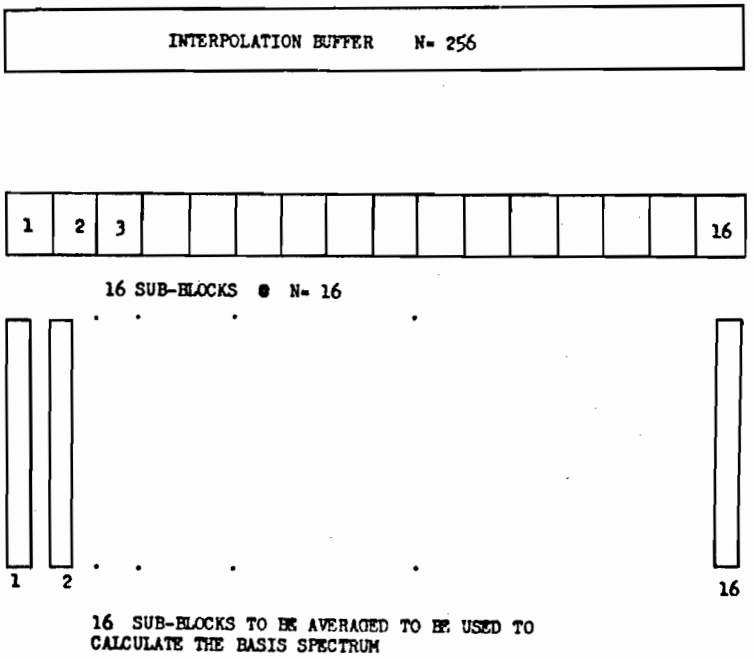
FIGURE 4-4

4.2.3 Side Information Interpolation

The computational complexity of an adaptive transform coder can be decreased by implementing the frequency domain transform using shorter frame lengths. Figure 2-3 illustrates the gain over PCM obtainable with transform coding using various block lengths. A multifold decrease in the computational complexity of the transform is accompanied by only a 2 dB drop in SNR. The main issue is, decreasing frame length increases frame rate making it impossible to transmit all the side information with the same frequency and precision for each frame. Decreasing the precision or frequency with which side information parameters are transmitted will permit side information to be compressed into the necessary bandwidth. Decreasing parameter precision by introducing quantization coarseness is not an attactive option. The first side information interpolation scheme investigated decreased the frequency of parameter update. Not every parameter needs to be encoded with the same accuracy or frequency. The scheme interpolates side information and permits each parameter to be updated at any desired frequency and accuracy. The number of bits available for side information each frame is distributed among the parameters by updating those parameters whose desired update frequency exceeds its actual update frequency. The parameters are quantized with the necessary precision until the bits available are exhausted. The scheme uses a look-up table containing update frequencies, desired update frequencies, and the number of quantization levels with which each parameter is quantized. The list of desired freqencies is searched and, if the desired frequency is greater than its occurrence, and if sufficient bits remain, that coefficient is updated and the list continues to be searched. If, at the end of the process, bits

remain unassigned, parameters not updated, are updated on a priority basis according to the desired frequency.

Another side information scheme was suggested by Zelinski and Noll [4]. The algorithm reduces the precision and frequency with which side information parameters are updated. The scheme exploits the spatial correlation of the parameters instead of the temporal correlation. Figure 4-5 shows an analysis frame of 256 samples divided into 16 sub-blocks. Each sub-block is transformed separately and the transform variances are averaged over the sub-blocks and this averaged variance is used in basis spectrum formulation. The sub-blocks are encoded and transmitted after the basis spectrum has been calculated and used for step size adaptation and bit assignment.



SIDE INFORMATION INTERPOLATION

FIGURE 4-5

4.2.4 Side Information Parameter Statistics And Quantization

Section 4.2.2 presents the SNR and segmental SNR performance of the LPC adaptive transform coder for various frame lengths.The effects of side information interpolation and quantization are also presented. Figure 4-4 shows rapidly declining SNR values for decreasing frame sizes when the side information is quantized. The explanation for this result can be obtained by investigating side information parameter statistics.

Side information contains all data required to reconstruct the basis spectrum at the receiver. A number of reflection coefficients, pitch period, pitch gain and an average energy parameter are needed to achieve this result. Side information parameter quantizers have predetermined and fixed characteristics. The results given in Section 4.2.2 ultilize side information quantizers with the same characteristics for every frame length. For some parameters this is inappropriate.

Figures 4-6 show histograms of the reflection coefficients of the utterance, "The box of lemons was thrown beside the parked truck." DOUG7. The left of Figure 4-6 shows the resulting reflection coefficients for a block length of 16 samples. The right of this figure shows the resulting reflection coefficients for a block length of 256 samples. Comparing these, it can be seen that the statistics of the reflection coefficients change only slightly, and therefore, there is marginal value in adjusting the reflection coefficient quantizers.

The pitch and pitch gain parameters lose significance for frame lengths of 16 and 32 samples. The explaination for this is simple. For short voiced speech segments, at most, one pitch period is contained within a single frame. This is an insufficient time period to gain an accurate estimate of the pitch. For lower frame lengths, it is not possible to exploit the perceptual aspects of pitch.

The average energy parameter is in fact the energy of the residual of the LPC spectral fit. This parameter linearly scales amplitude of the basis spectrum. It therefore has no effect on bit assignment but, has a significant effect on step size of the transform coefficient quantizers. Energy of the residual is dependent on the frame length via two mechanisms. The length of the frame directly effects the energy of the residual because the frame length determines the limits over which the residual is summed. Therefore, the residual energy for a frame length of 256 samples is expected to be 16 times larger than the energy for a frame length of 16 samples. Figure 4-7a and 4-7b are plots of the average energy parameter, for 16 and 256 sample frame lengths. This parameter is 10 times smaller for a block length of 16 samples as is shown in figure 4-7. The reason why the parameter is twice as large as expected is because the normalized residual energy is larger for shorter block lengths. The ability of the LPC analysis to obtain a good spectral fit to the variance of the transform coefficients declines with frame length causing an increasing residual energy.

Careful quantization of the side information parameters is essential for maximum coder performance because coder performance is strongly effected by the accuracy of the short term spectral estimate.

The reflection coefficients are quantized using a quantizer employing a log-area ratio non-linearity. The reason for using this non-linearity is because deviations in log-area quantized reflection coefficients results in uniform spectral pertubations. [17] Uniform quantization of reflection coefficients results in non-uniform spectral pertubations. The minimiun and maximum break points of the log-area quantizer are listed in table 2 with the corresponding number of quantization levels. The reflection coefficients quantized in this manner are easily transmited within 1.5 khz of bandwidth. Fewer quantization levels may be employed if the bits are required to combat perceptual distortions such as frame boundry discontinuities (see section 4.2.6).

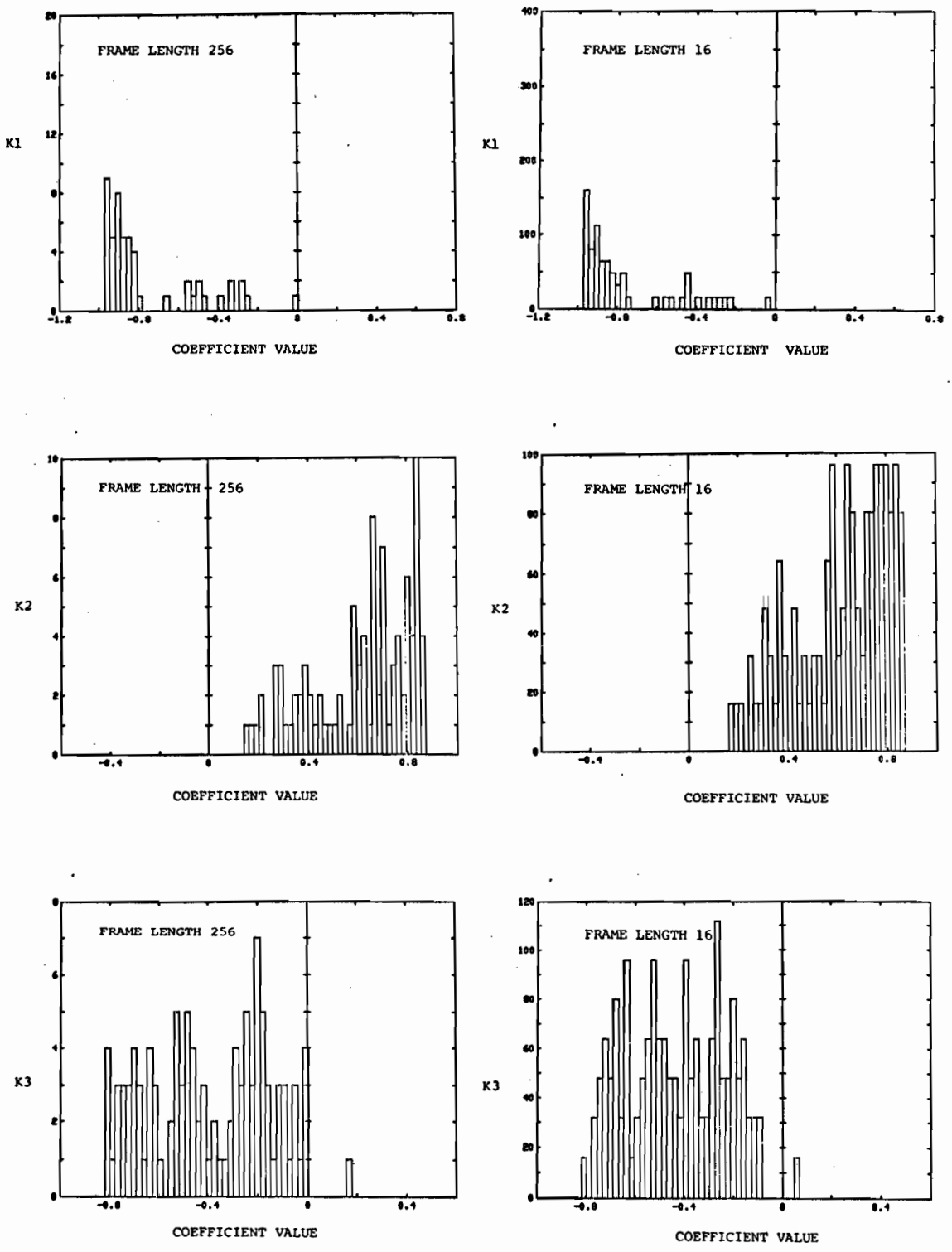
TABLE 2

REFLECTION COEFFICIENT QUANTIZER SPECIFICATIONS

reflection coefficient	minimum breakpoint value	maximum breakpoint value	number of levels
K1	-0.97	0.77	64
K2	-0.37	0.89	64
K3	-0.82	0.57	64
K4	-0.30	0.76	64
K5	-0.47	0.38	64
K6	-0.20	0.64	64
K7	-0.26	0.41	32
K8	-0.17	0.58	32

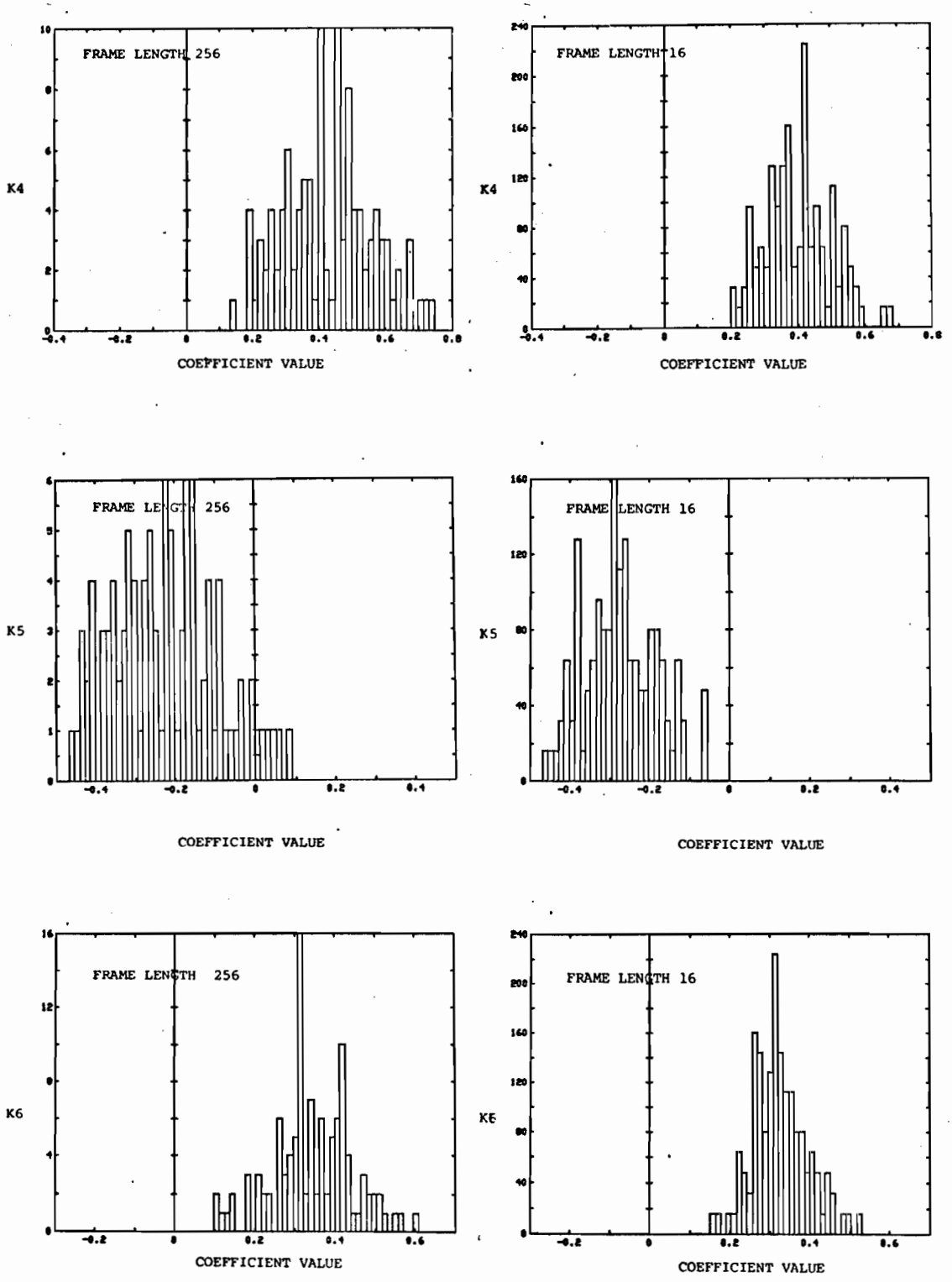
In summary, for frame lengths of 16 and 32 samples, the perceptual aspects of pitch cannot be exploited. The reflection coefficient statistics do vary with frame length but not significantly enough to effect SNR figures drastically. The average energy parameter varies greatly with frame length and requires quantization with quantizers with a characteristic adjusted to each frame length.

Figures 4-8a to 4-8h show the SNR and segmental SNR for DOUG1 to DOUG8 when the average energy parameter is adjusted for each frame length. Compare these with Figure 4-4.



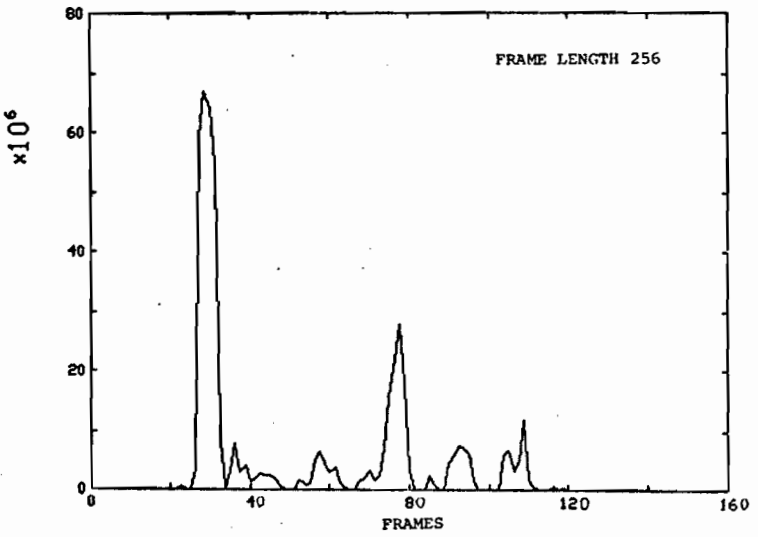
REFLECTION COEFFICIENT HISTOGRAMS

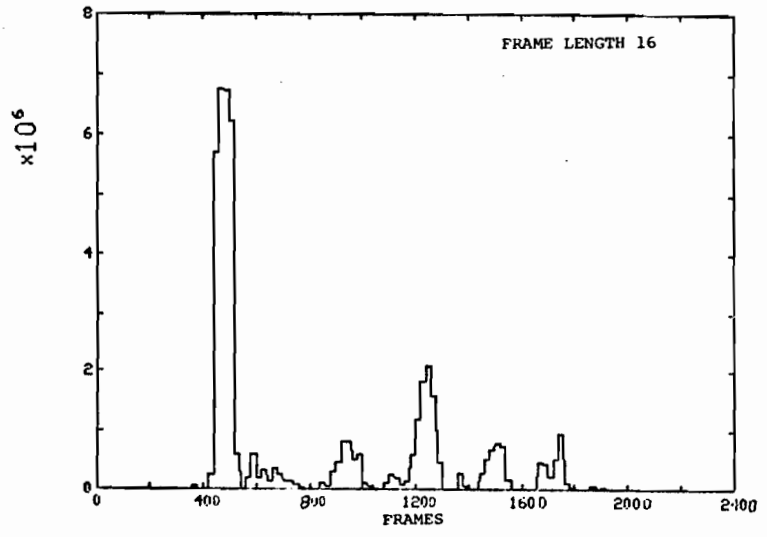
FIGURE 4-6



REFLECTION COEFFICIENT HISTOGRAMS

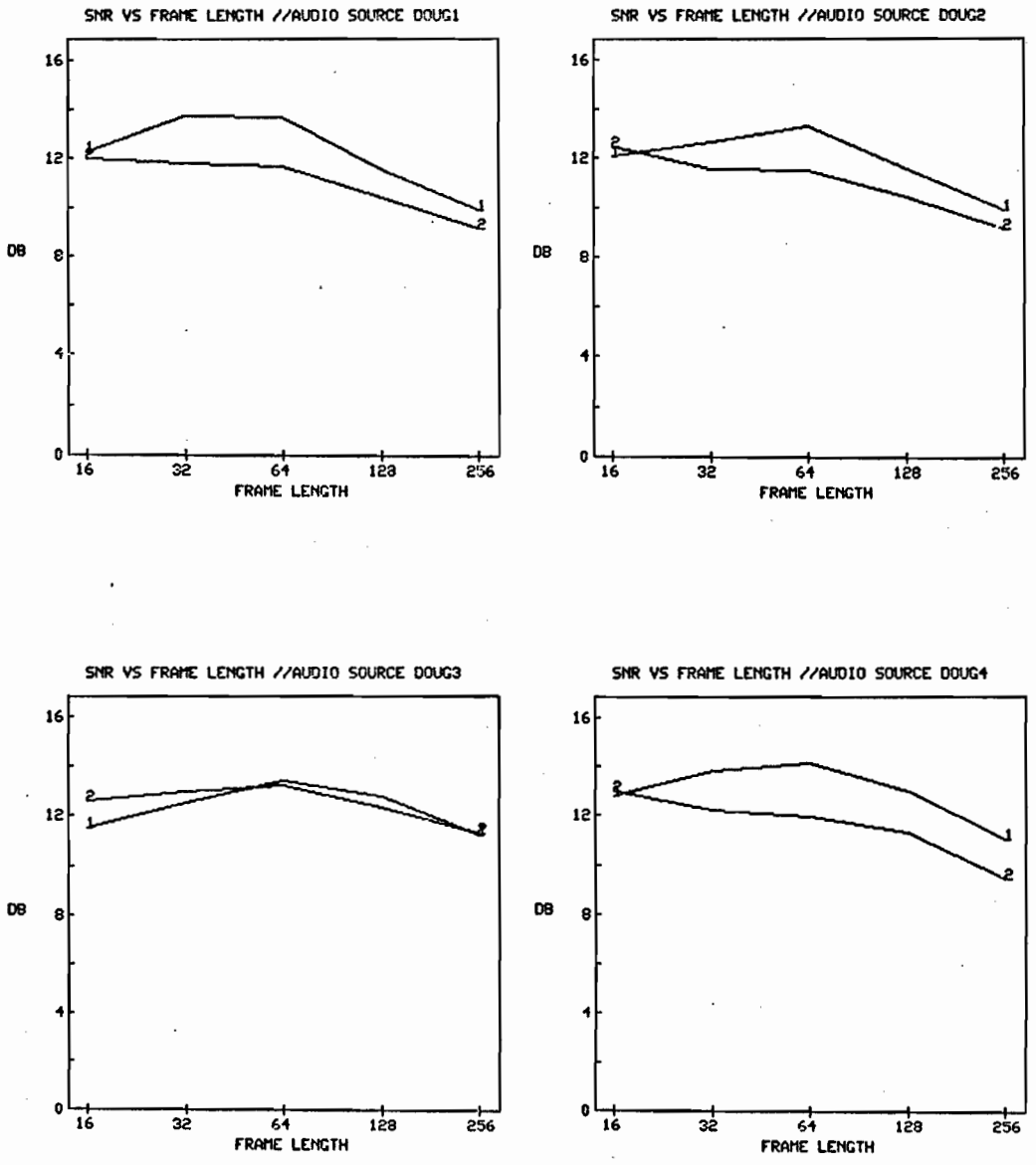
FIGURE 4-6





AVERAGE ENERGY PARAMETER

FIGURE 4-7

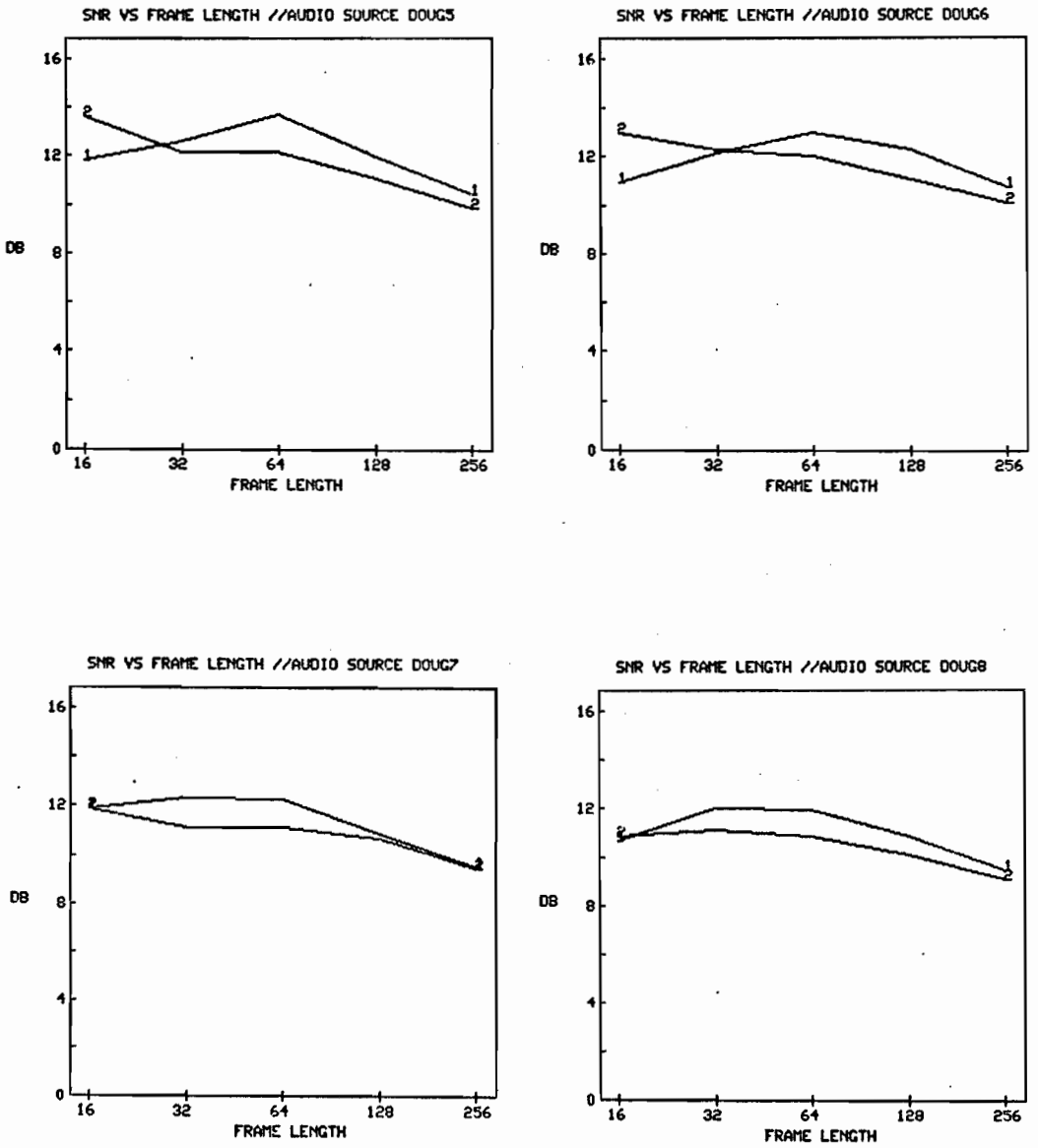


QUANTIZED SIDE INFORMATION
9.6 KB/SEC ATC CODER

1 SNR INTERPOLATING
2 SEG. SNR INTERPOLATING

CODER SNR PERFORMANCE

FIGURE 4-8



QUANTIZED SIDE INFORMATION
9.6 KB/SEC ATC CODER

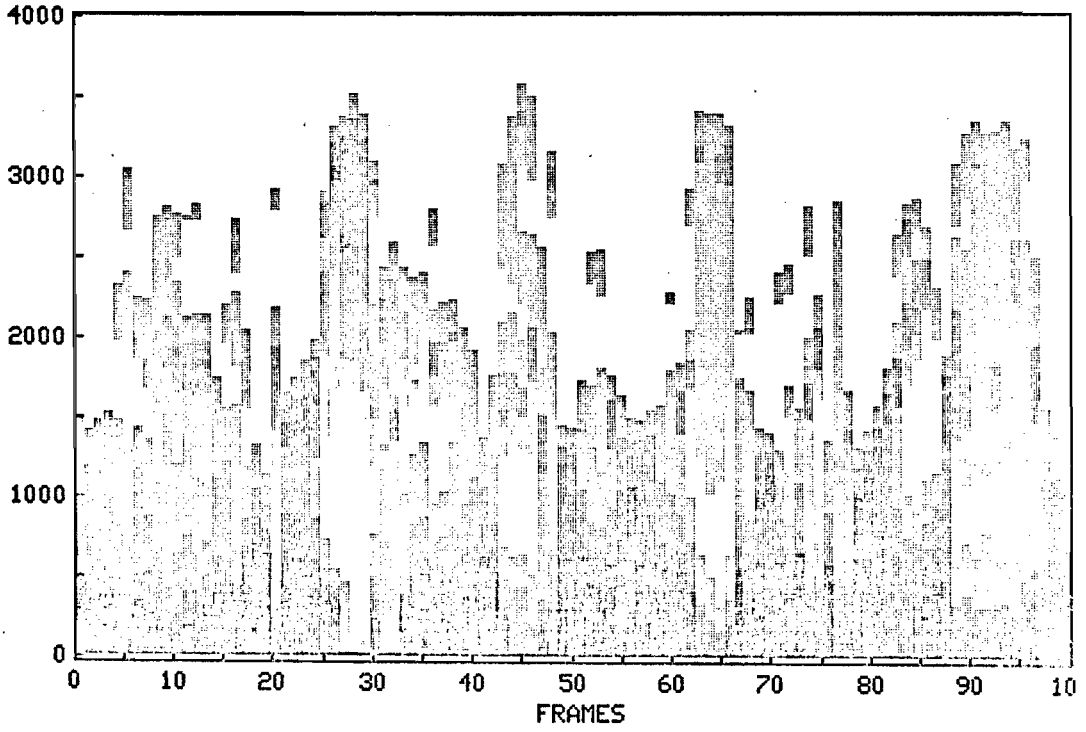
1 SNR INTERPOLATING
2 SEG. SNR INTERPOLATING

CODER SNR PERFORMANCE

FIGURE 4-8

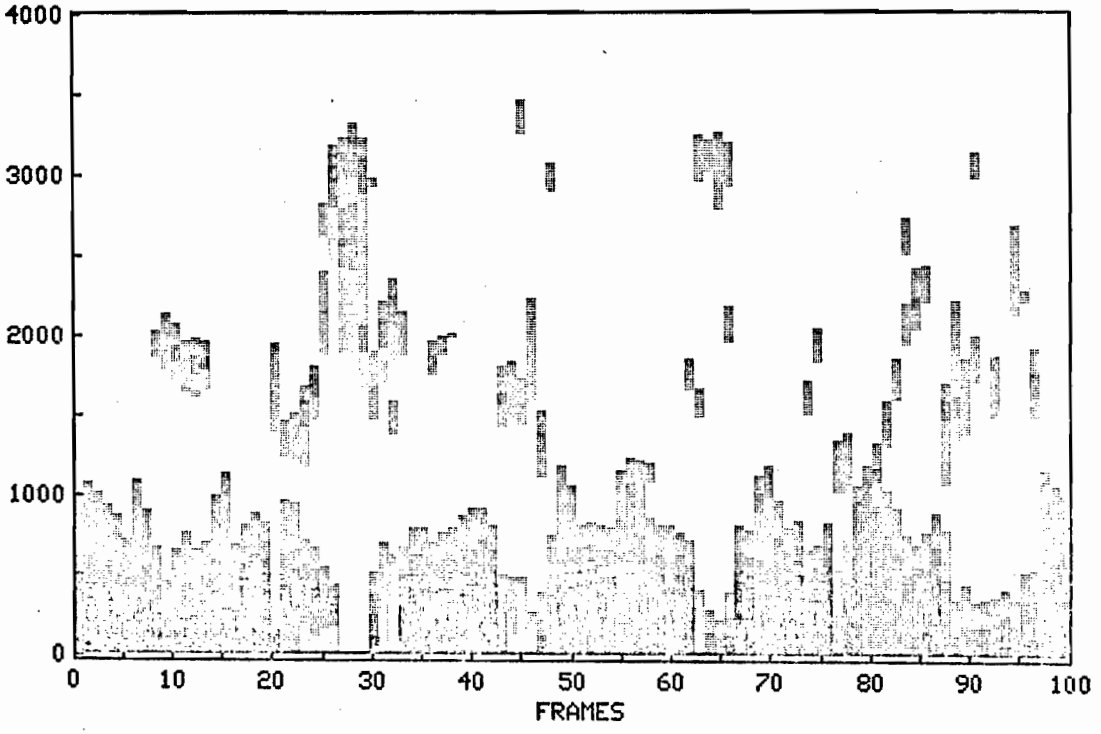
4.2.5 The Low-Pass Effect

Not all transform coefficients can be transmitted at bit rates of 16 kb/sec and below. The optimal bit assignment discards those transform coefficients whose variance is below the average distortion. As a result the low level portions of the spectrum at high frequencies are discarded. This leads to an audible reduction in the transmitted bandwidth generaly refered to as the low-pass effect. In other words, the SNR maximimizing bit assignment rule leads to excessive concentration of bits in the formant frequencies. Figure 4-9a and 4-9b show the visable bit assignment for a frame length of 256 at (16 and 9.6 kb/sec respectevely) samples for the sentence DOUG1. The darkness of a region is an indication of the number of bits allocated at the time to a particular transform coefficient. A concentration of bits at low frequencies can clearly be seen for the 9.6 kb/sec coder. Many approaches for solving the low-pass effect have been suggested by Zelinski and Noll [4]. Substitution of the non-transmitted transform coefficients with white noise was adopted. The zero mean unit variance white noise was scaled by the basis spectrum so that its variance would match the variance of the transform coefficients. This technique was found to add some bandwidth to the adaptive transform coder. However for frame lengths of 16,32 and 64 samples frame boundary discontinuitities distortion is more significant than the low-pass effect.



VISIBLE BIT ASSIGNMENT

FIGURE 4-9.A



VISIBLE BIT ASSIGNMENT

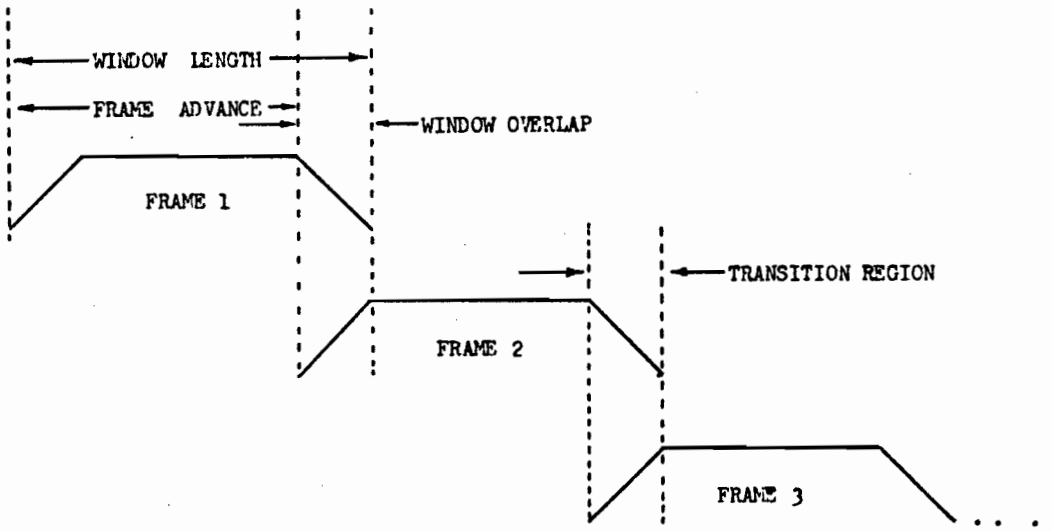
FIGURE 4-9.B

4.2.6 Frame Boundary Discontinuities

The transform coding scheme discussed previously does not guarantee signal continuity at the frame boundaries. The effect of a discontinuity is to spread signal energy out in the frequency domain. Perceptually, a frame discontinuity is heard as a sharp click. Frame boundaries occur at regular intervals. The perceived noise frequency depends on the interval between frame boundaries. As the frame length shrinks the discontinuities become more frequent and audible. The dominant impairment of speech quality for frame lengths of 128 and 256 samples is due to the low-pass effect. Frame boundary discontinuities are perceptible but do not detract significantly from speech quality. For frame lengths of 16 and 32 samples some method of reducing frame boundary discontinuity distortion is required.

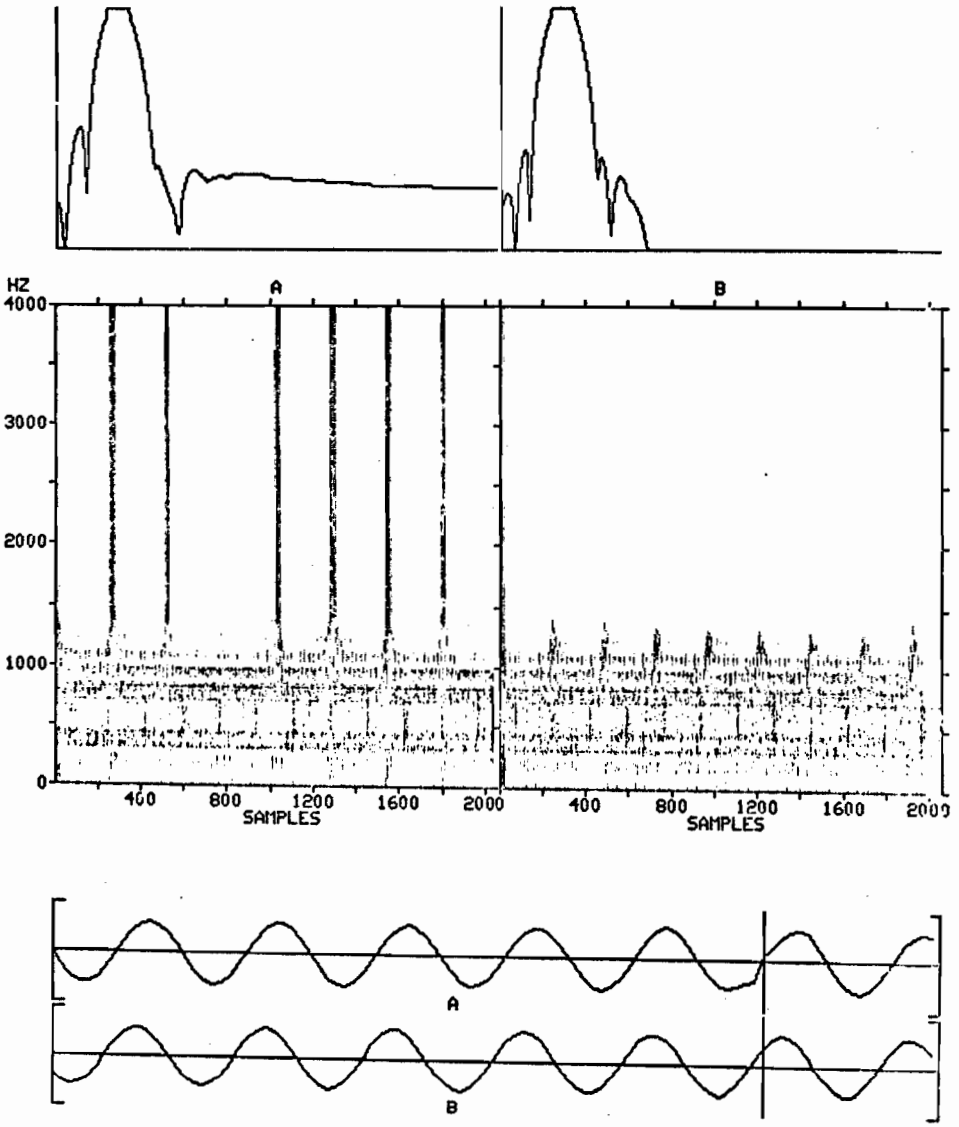
The technique employed to reduce these frame boundary discontinuities is by the use of a raised cosine window. The input speech is processed in overlapping windowed frames. After each frame is coded, the frame is advanced by less than the window length. The next analysis frame is calculated such that there is a transition region where the windows are overlapped. Adding coded speech from overlapping frame portions ensures a smooth transition across frame boundaries. A further improvement is possible by windowing the coded speech rather than the input speech. Continuity at the frame boundaries is not corrupted by quantization error if the coded speech is windowed. This procedure is illustrated in Figure 4-10. The results of this procedure for a 550 Hz sinusoid coded at 9.6 kb/sec using a frame length of 256 samples is shown in Figure 4-11. The spectrogram clearly displays a reduction of frame boundary

discontinuity distortion when a raised cosine window is employed. A comparison of waveforms is shown below. Waveform B illustrates how windowing reduces discontinuities. The profile of the energy distribution, and the spectrogram are shown at the frame boundries showing how discontinuities introduce energy into the high frequency portions of the spectrum. Table 3 (p. 4-40) presents frame advance values which give good results for various frame lengths.



ANALYSIS FRAME WINDOWING

FIGURE 4-10



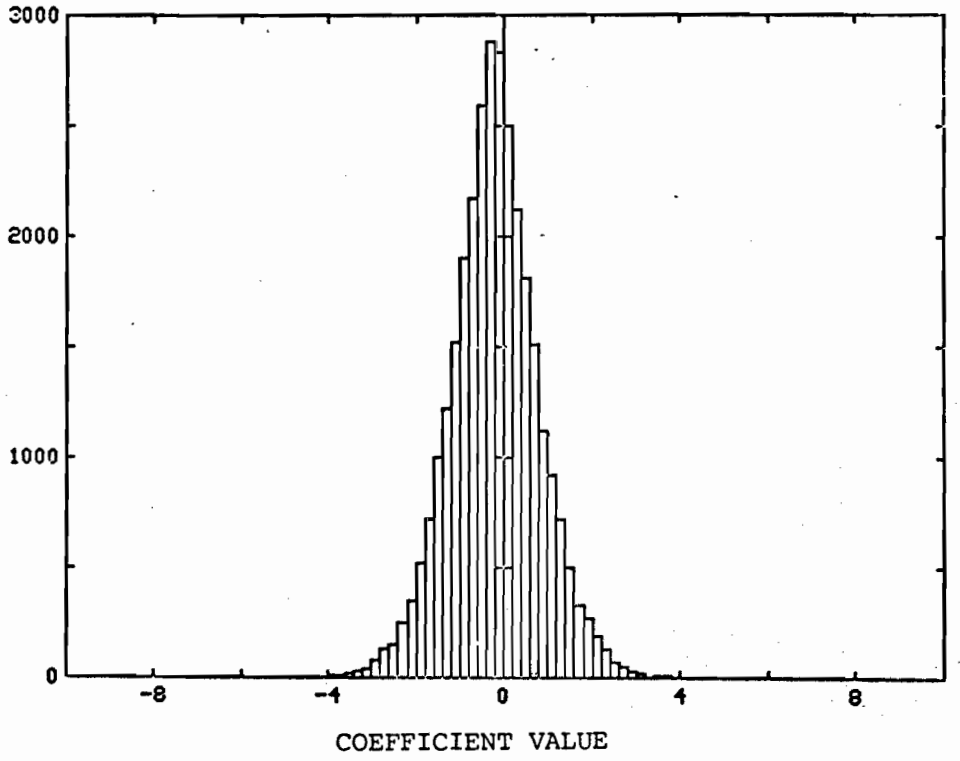
FRAME BOUNDARY DISCONTINUITY REDUCTION

FIGURE 4-11

4.2.7 Transform Coefficient Statistics

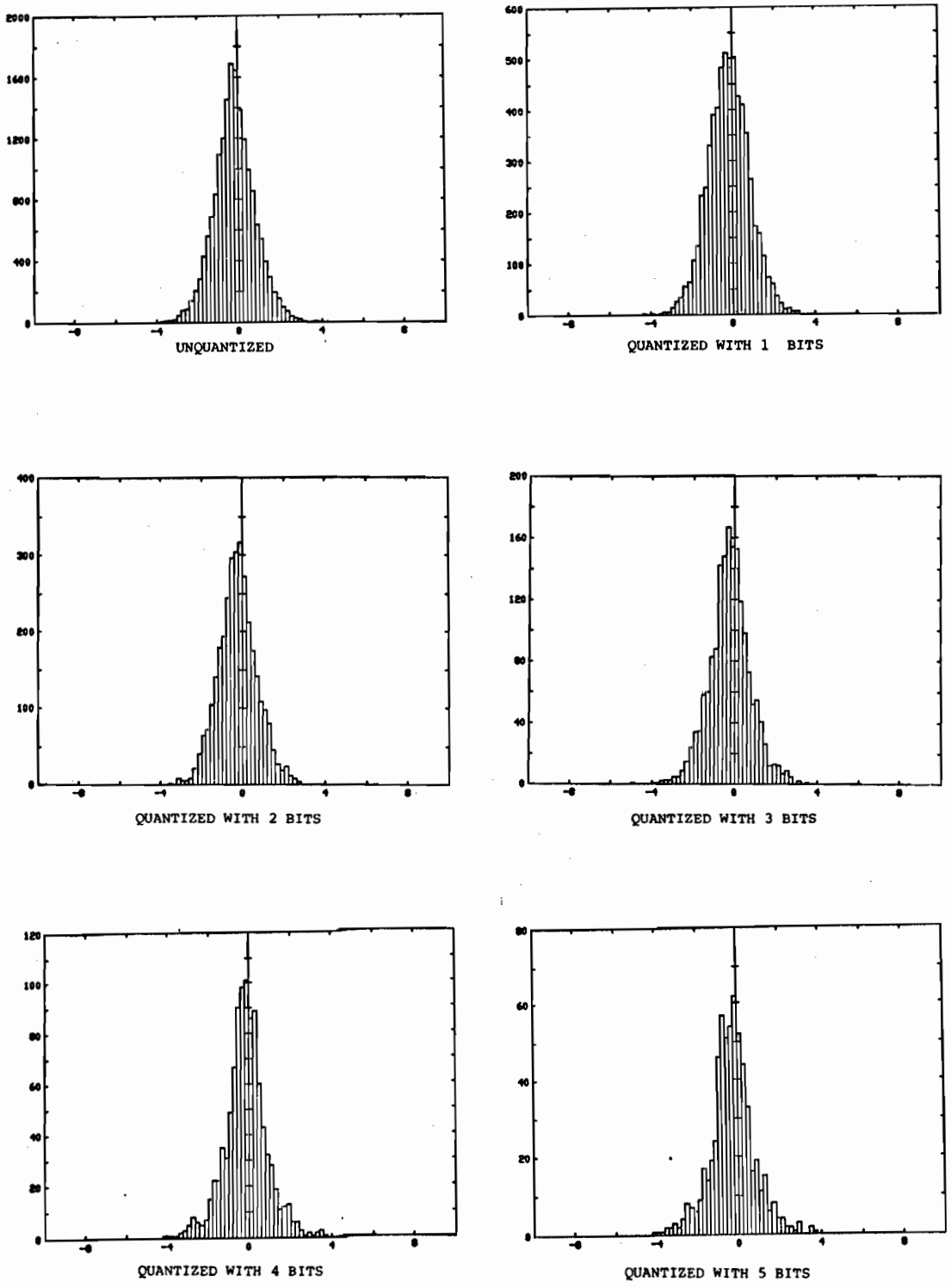
The normalized transform coefficient distribution is illustrated in Figure 4-12 for the sentence DOUG1. The distribution is very nearly Gaussian with zero mean and unit variance. This indicates that the pre-scaling effect discussed in Section 2.2 does not significently distort the transform coefficient distribution. The normalized transform coefficients can be divided into classes according to the number of quantization bits assigned to each coefficient. These classes are illustrated in Figure 4-13 and are all aproximately Gaussian with zero mean and unit variance. Recall the basis restricted quantization approach assumed that the transform coefficient distribution was Gaussian.

In Figure 4-14, the performance of optimal (mean square error) quantizers are compared to Gaussian quantizers for 2,3,4 and 5 bits. The optimal quantizers are tailored to the statistics of the sentence to achieve maximum SNR. The Gaussian quantizer is optimal only for Gaussian input signals. This figure illustrates the effect on SNR performance of varing the input signal level. The results reinforce the contention that transform coefficients are Gaussian. There seems little to be gained by designing an optimal quantizer for the long term statistics of speech and speakers.



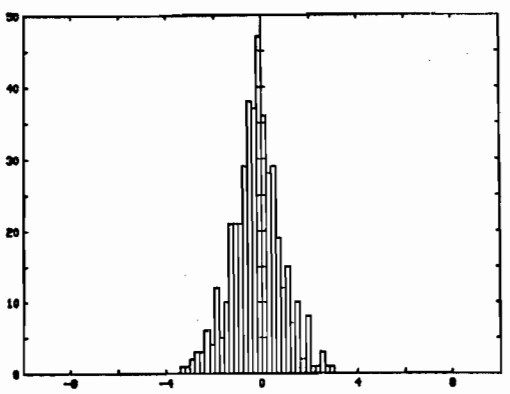
TRANSFORM COEFFICIENT HISTOGRAMS

FIGURE 4-12

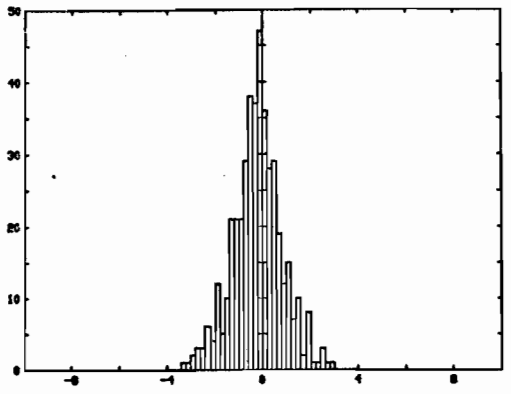


TRANSFORM COEFFICIENT HISTOGRAMS

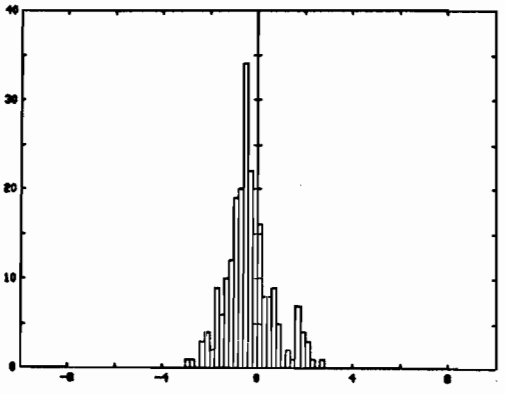
FIGURE 4-13



QUANTIZED WITH 6 BITS



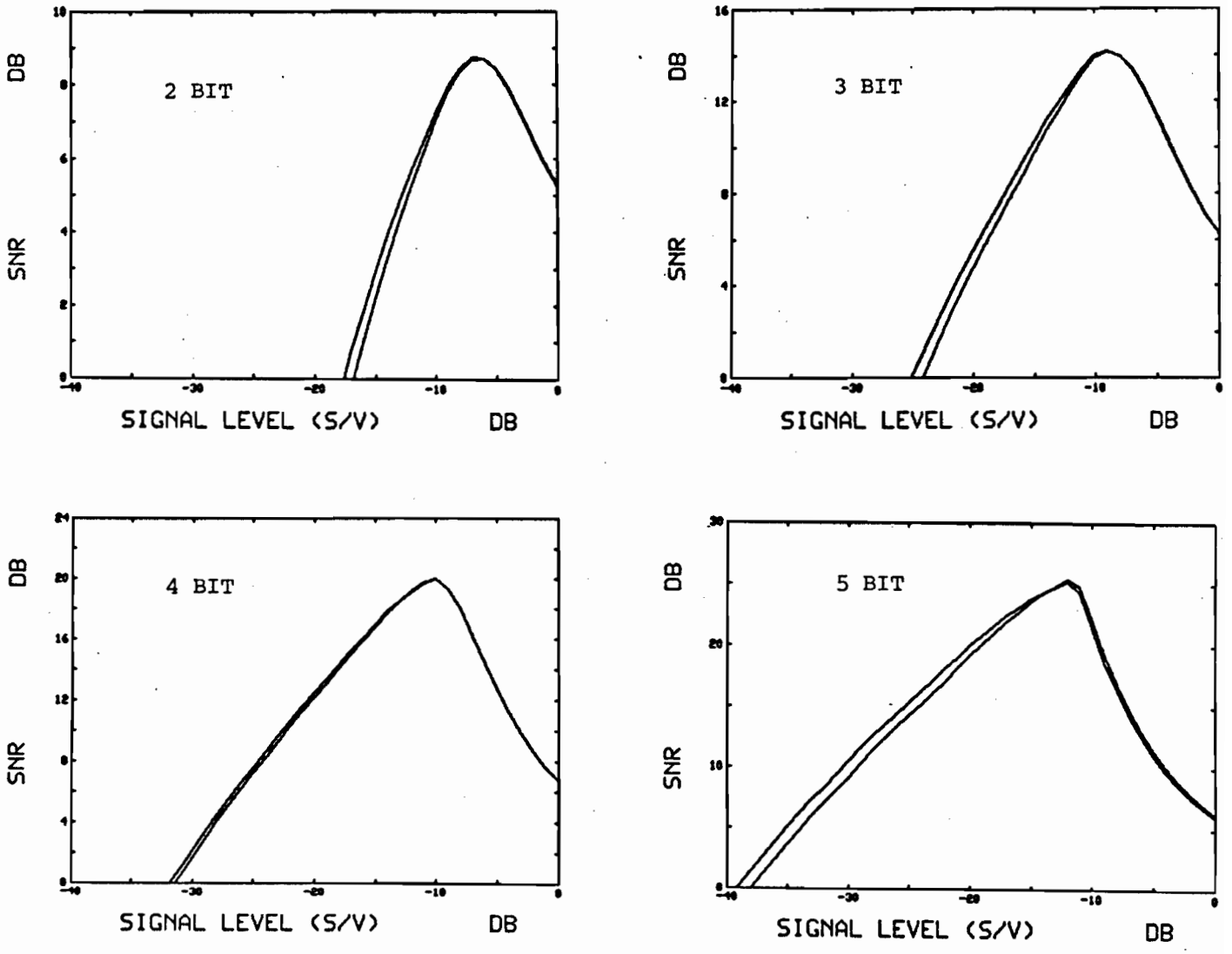
QUANTIZED WITH 7 BITS



QUANTIZED WITH 8 BITS

TRANSFORM COEFFICIENT HISTOGRAMS

FIGURE 4-13



S standard deviation of the input signal
V maximum quantized level

TRANSFORM COEFFICIENT QUANTIZER PERFORMANCE

FIGURE 4-14

4.2.8 Subjective Effect Of Pre-emphasis And Spectral Shaping

Pre-emphasis of the input signal results in better high frequency reproduction by boosting higher frequencies before coding. More bits are then assigned to high frequencies and these components are reproduced more accurately at the receiver. De-emphasis returns the high frequency components back to their original levels. De-emphasis reduces frame boundary discontinuity noise by attenuating the high frequency components of this noise. Unfortunately too much pre-emphasis causes signal distortions by assigning too large a proportion of bits to the higher frequencies. Pre-emphasis is accomplished by forming the first difference of the input signal with the past sample, weighted by a pre-emphasis factor ζ . That is

$$Y(I)=X(I)-\zeta\ X(I-1)$$

Bit allocation determines the accuracy with which transform coefficients are quantized and thus the quantization noise distribution. A necessary condition for minimum coding error (MSE) is that all transform coefficients suffer from the same amount of distortion. Therefore, a bit assignment rule based on minimum mean square error over the analysis block, leads to a flat noise distribution. A flat noise distribution is not the most perceptually desirable. Noise should be shaped to take advantage of auditory masking in the human hearing mechanism. This can be done by multiplying the basis spectrum by a shaping function. This in turn effects the bit assignment. Pitch is added after the basis spectrum is warped and is not effected by the shaping. The shaping function selected is the basis function raised to the power γ. As the value of is varied between $-1<\gamma<0$ the noise spectrum evolves from one which

following the speech spectrum to that of a flat distribution. The effect of this parameter is to change the peak to valley ratio of the basis spectrum. Flattening the basis spectrum results in more bits assigned to the high frequency components. In the extreme, a value for γ of -1 results in a flat warped basis spectrum and a constant bit assignment. With each transform coefficient quantized with the same number of bits the quantization noise follows the speech spectrum. Values for the pre-emphasis factor and the spectral shaping parameter which yield good speech reproduction are shown in Table 3.

TABLE 3

FRAME ADVANCE AND SUBJECTIVE PARAMETER VALUES

Frame Length	Frame advance	Pre-emphasis	Spectral Shaping factor
256	240	.1	-0.1
128	120	.1	-0.1
64	56	.1	-0.1
32	24	.1	-0.1

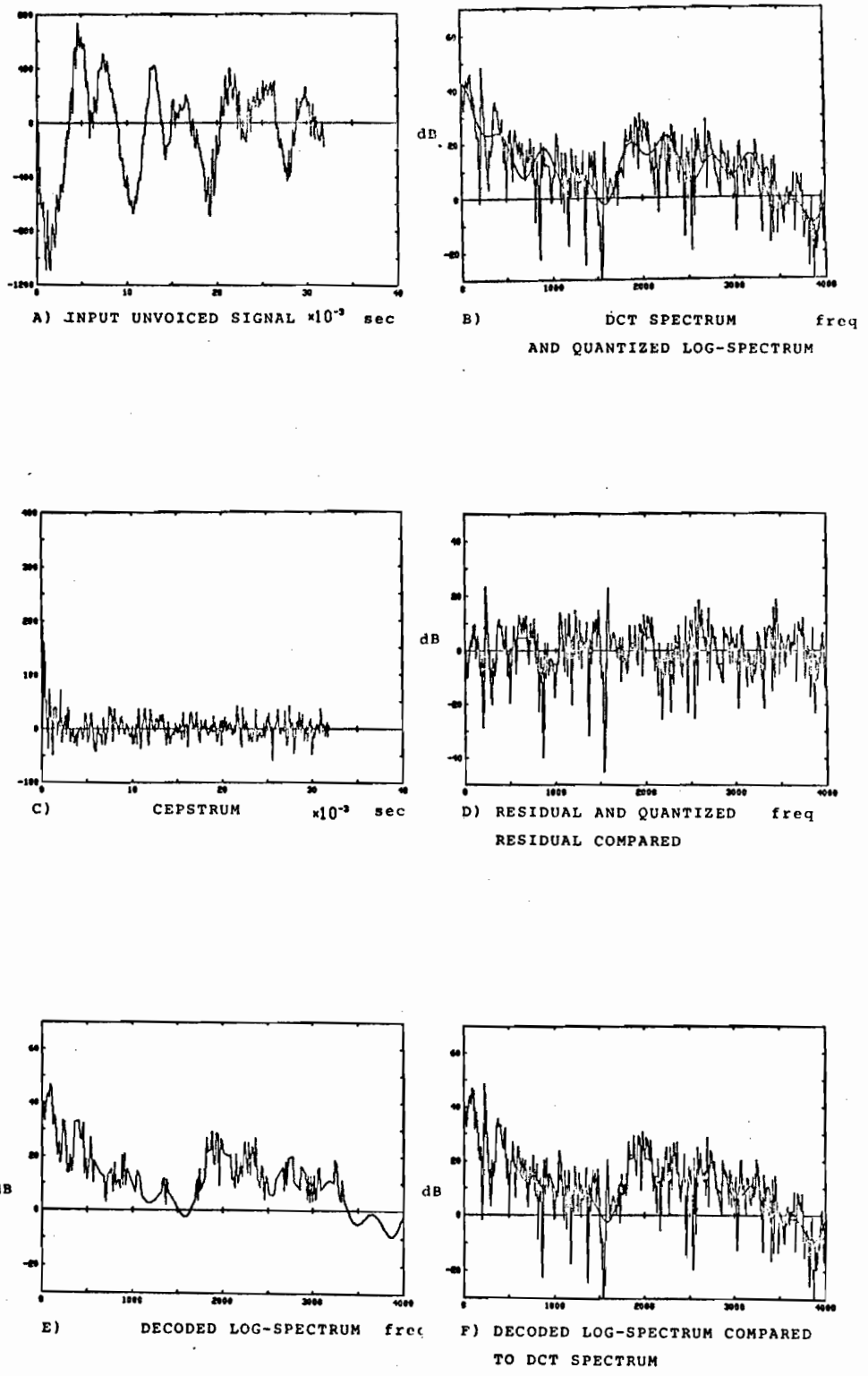
4.3 THE HOMOMORPHIC CODER

In this section the operation and performance of the homomorphic adaptive transform coder is discussed. It is shown that the log-spectrum output of the homomorphic model is in a convenient form for bit assignment and for transform coefficient quantization. The effects of reduced frame size, parameter quantization and side information interpolation are presented. Analysis frame windowing effects and their ability to reduce frame boundary discontinuities are studied. The homomorphic and LPC adaptive transform coders are compared.

4.3.1 Coder Operation

The Homomorphic Adaptive Transform Coder structure is reviewed in Section 3.3. A block diagram of the coder structure is presented in Figure 4-9. Each waveform in the coding process is illustrated in Figure 4-15. A typical 256 point unvoiced speech frame is shown in Figure 4-15a. In this coding scheme the DCT spectrum is considered to be an estimate of the spectral variance of the transform coefficients. The log-spectrum is a smooth fit to the logarithm of the DCT of a speech frame. Both are shown in Figure 4-15b. The log-spectrum is obtained from a homomorphic analysis as follows. The inverse discrete cosine transform of the logarithm of the DCT of a speech frame is called the pseudo-cepstrum and is shown in Figure 4-15c. The low-time region of the pseudo-cepstrum is windowed and the pitch peaks are isolated. The DCT of the windowed cepstrum with added pitch peaks is the log-spectrum. The quantized log-spectrum is the DCT of the quantized cepstrum and analagous to the basis spectrum of the LPC adaptive transform coder. Step size adaptation and bit assignment are based on the quantized log-spectrum. The quantized pseudo-cepstrum is a parameterized estimate of the variance of the transform coefficients and is transmitted to the receiver with the pitch peaks as side information. Step size adaptation of the transform coefficients is accomplished indirectly. The residual of the logarithm of the DCT and the quantized log-spectrum is quantized and is transmitted to the receiver. Quantizing the residual in the log domain is equivalent to quantizing the ratio of the transform coefficients and spectral estimate in the linear domain. This is the mechanism by which step size adaptation is accomplished in the LPC coder. The residual and quantized residual are shown in Figure 4-15d. Additional side

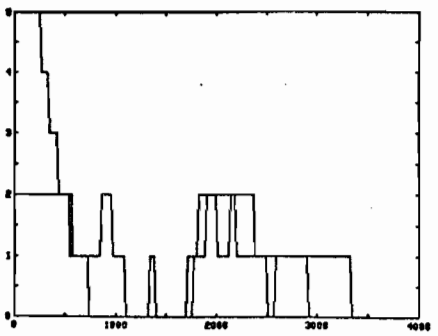
information namely the sign of the transform coefficients are required by the receiver. The quantized log-spectrum obtained from the quantized cepstrum parameters is added to the residual at the receiver to yield a decoded log-spectrum shown in Figure 4-15e. The quantized DCT and unquantized DCT are compared in Figure 4-15f. Absence of the low-pass effect in this figure is because the signs of the transform coefficients were transmitted with the side information. In effect, the minimal value of the bit assignments is one. An algorithm was also developed to permit zero bit assignments. The algorithm described in Section 2.3 is modified so a bit assignment of one indicates that only the sign of the transform coefficient is sent. This complicates some of the decoders illustrated in Figure 3-6. Figure 4-15g is a comparison of bit assignments for the LPC and homomorphic coders. Figure 4-15h presents a comparison of the speech frame and coded speech frame. Figure 4-15.i compares the basis spectrums for the LPC and homomorphic analysis.



HOMOMORPHIC ADAPTIVE TRANSFORM CODER

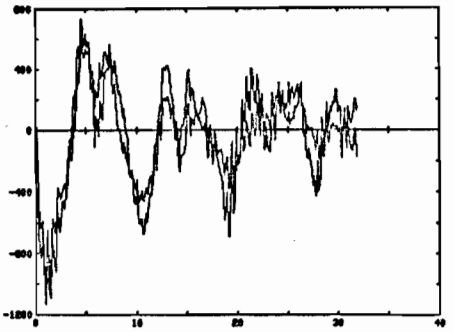
WAVEFORMS

FIGURE 4-15



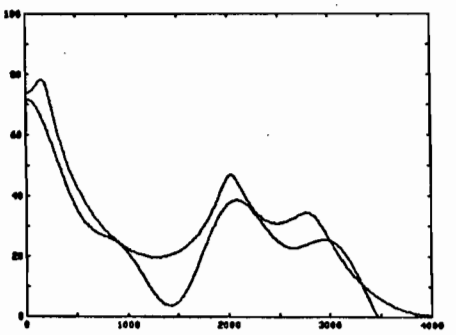
G) BIT ASSIGNMENT FOR LPC AND HOMOMORPHIC CODERS

freq



H) INPUT AND OUTPUT SIGNALS

$\times 10^{-3}$ sec



dB

freq

I) BASIS SPECTRUM AND QUANTIZED LOG-SPECTRUM COMPARED

HOMOMORPHIC ADAPTIVE TRANSFORM CODER WAVEFORMS

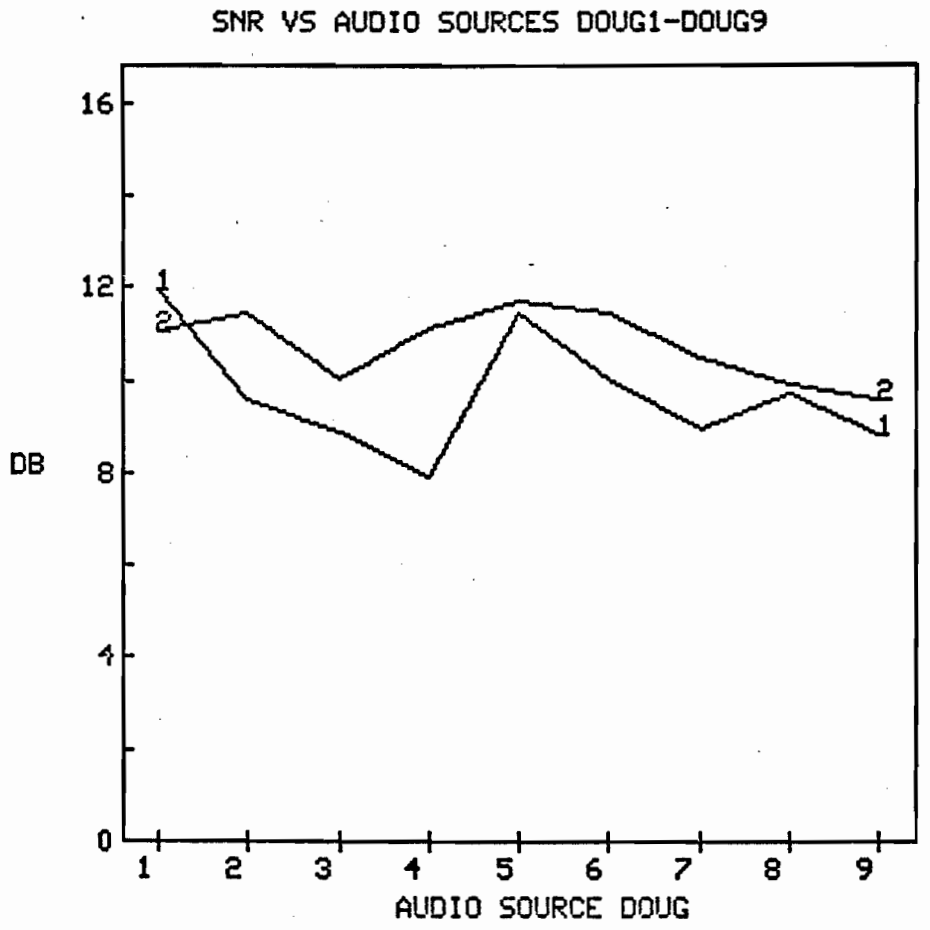
FIGURE 4-15

4.3.2 Coder Performance

The homomorphic coder implemented, operating at 9.6 kb/sec with a frame length of 256 samples, generates speech with an average segmental SNR of 11 dB. Figure 4-16 shows the SNR and segmental SNR for nine sentences. Cox and Crochiere [8] report segmental SNR performance of 13 dB. The homomorphic coder does not concentrate bits in the formant regions of the spectrum. This is thought to be the reason why the homomorphic coded speech has less perceptual low-pass distortions than the LPC coder. Homomorphic coding introduces frame boundary discontinuities similar in nature to distortions introduced by LPC coded speech. The raised cosine windowing is effective in alleviating this problem.

The homomorphic coder is capable of coding speech using shorter frame lengths. The approach is similar to the LPC coder. An averaged DCT spectrum is generated and the homomorphic side information processing is performed on the averaged spectrum. The quantized log-spectrum and thus the bit assignment will remain the same for as many sub-blocks as are required to add up to 256 samples. Figure 4-17 shows for various frame lengths the SNR and segmental SNR performance of the coder for the sentence DOUG1.

In short the homomorphic coder produces results comparable with the LPC coder using a radically different coder structure. Perceptual coded speech quality is better for the homomorphic speech coder despite it's poorer SNR performance. This is sufficient reason to warrant its investigation.

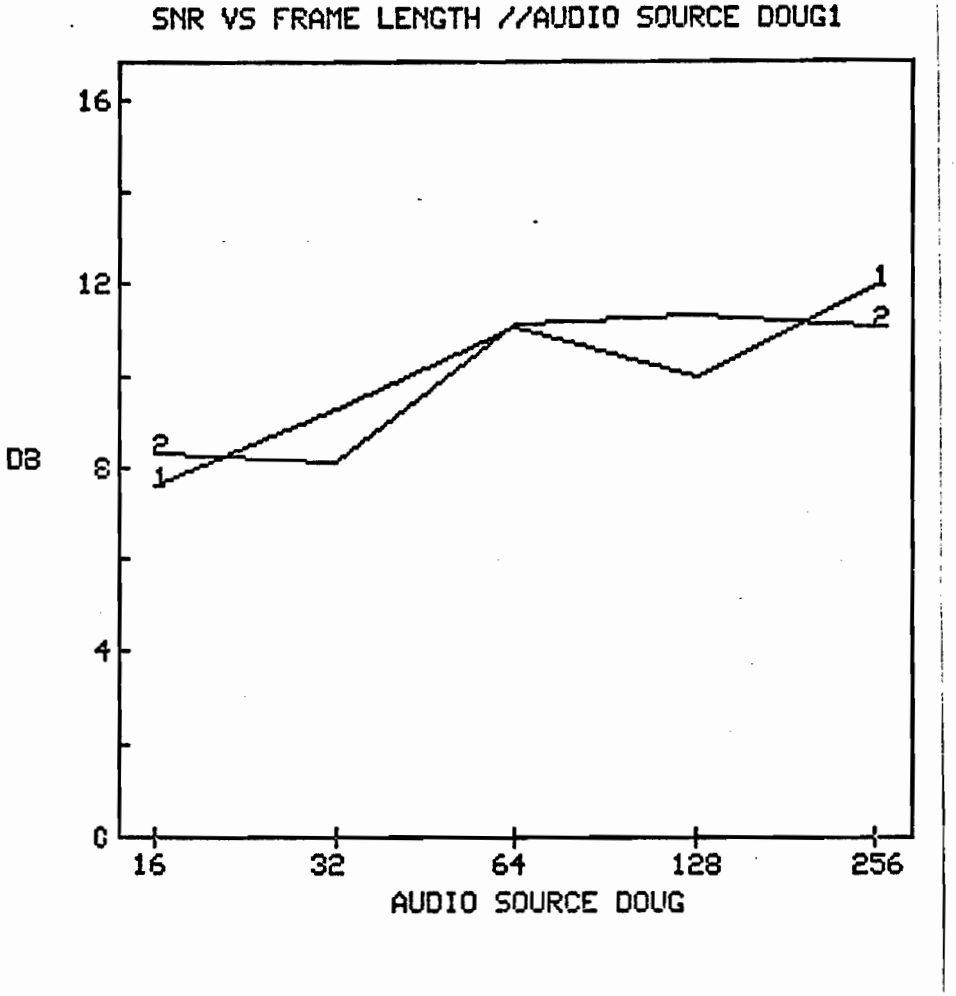


UNQUANTIZED SIDE INFORMATION
9.6 KB/SEC ATC CODER

1 SNR
2 SEG. SNR

CODER SNR PERFORMANCE

FIGURE 4-16



UNQUANTIZED SIDE INFORMATION
9.6 KB/SEC ATC CODER

1 SNR INTERPOLATING
2 SEG. SNR INTERPOLATING

CODER SNR PERFORMANCE

FIGURE 4-17

CHAPTER 5

CONCLUSIONS

The basic structure of transform coding has been presented (in Chapter 2) with justification for selection of the mean square error criteria. A basis restricted quantization strategy has been selected and optimal bit assignment algorithms have been discussed. Optimal and sub-optimal transforms have been introduced and compared. The discrete cosine transform was selected for use by the coders for practical reasons. The theoretical discussion was subsequently expanded (in Chapter 3) to include the practical problems of adaptive transform coding and considers three spectral parameterization techniques. Later (in Chapter 4), two coders using all-pole and homomorphic modelling of the short term spectrum were evaluated. In this connection the use of reduced frame lengths and side information interpolation schemes were discused as a method of decreasing coder computational complexity. Frame boundary discontinuities and low-pass filtering effects were presented as the primary sources of perceptual distortion. Techniques were evaluated for reduction of these distortions. Subjective effects of pre-emphasis and spectral shaping are also discussed. The homomorphic and LPC coders have radically different structures but offer aproximately the same speech quality.

There is a growing conviction among current investigators that the homomorphic structure may be more suited to real-time implementation [8].

Results from simulations of the LPC adaptive transform coder at 9.6 kb/sec indicate that there are three identifiable distortions. These are, in order of importance, frame boundary discontinuities, the low-pass effect and moving dead bands.

Frame boundary discontinuities are the most severe perceptual degradation in transform coding. They are perceived as a rapid clicking in the background and are also present during silent periods. Reducing the analysis frame length makes these discontinuities more frequent and disturbing. The problem is aggravated at low bit rates due to course quantization of the transform coefficients. In Section 4.2.6 overlapping raised cosine windows, designed to ensure a transition region between frames, were found to reduce the problem greatly. The overlapping of the analysis windows requires some source samples to be coded twice. This results in an increased overall bit rate. From a subjective speech quality point of view, bits used by windowing at the expense of the transform coefficient and side information quantizers are well spent.

At low bit rates transform coding only assigns bits to frequency regions with high energy content. The energy in speech declines with frequency and the coded speech suffers from a loss of bandwidth. The speech is, in effect, low-pass filtered because the coder is not assigning bits to the high frequency portions of the spectrum. The nature of this distortion is such that it does not disturb the listener or detract from the intelligibility of the speech. Further,

the distortion is unavoidable at low bit rates since there are insufficient bits to assign to all transform coefficients. The homomorphic coder suffers from less loss of bandwidth than does the LPC coder. Cox and Crochiere [8] attribute this to the tendency for the LPC coder to concentrate bits in formant regions. Pre-emphasis improves reproduction of the high frequency portions of the spectrum by boosting the high frequencies prior to coding so that more bits are assigned in the high frequency regions. The speech is later de-emphasized and this has an added benefit of reducing frame boundary discontinuity noise. The perceived bandwidth of coded speech can also be increased by adding white noise to the high frequency regions.

For larger frame lengths it is possible to have no bits assigned to transform coefficients in the mid-band regions. Burbling and slurring of the speech is the result. Filling dead regions with noise alleviates this somewhat, however, at low bit rates there are insufficient bits to assign to all frequencies.

Quantization noise shaping was investigated by shaping the basis spectrum to effect the bit assignment without effecting the step size adaptation. Noise shaping can be interpreted as a type of pre-emphasis because it results in more bits being assigned to high frequency portions of the spectrum.

Performance of adaptive transform coding depends greatly on the accuracy of the basis spectrum estimate. The investigation of complex side information parameterization schemes is therefore justified. Further, careful attention to quantization of the side information is desirable. Investigation of the side information statistics indicate that quantization of the average energy parameter requires special

attention.

Transform coding is among the most complex of the speech coders. Simplifying transform coding will facilitate real time implementation. Performance of adaptive transform coding is strongly dependent on the accuracy of the basis spectrum estimate. Therefore, simplifying the spectral parameterization scheme is not recommended. Shortening the analysis frame lengths reduces the complexity of the transform but increases the amount of side information. The side information interpolation scheme discussed in Section 4.2.3 performs well. Shortening the analysis frame length from 256 samples to 16 samples results in a decline in SNR of only 1 to 2 dB. The major distortion present in speech coded with shorter frame lengths is frame boundary discontinuities. These can be reduced by windowing, but a greater proportion of bits must be assigned to windowing when using shorter frame lengths.

The performance of both coders were evaluated over the full 0-4000 Hz frequency range. Many authors evaluate their coders over the more restricted 300-3200 Hz frequency range. SNR performance figures must be interpreted accordingly. Performance tests also indicate that a 2 dB SNR improvement can be expected if two more reflection coeffients are transmited to the receiver.

Adaptive transform coding is a promising technique for speech coding at low to medium bit rates. There remains however, much more work to be done in this field. The effect of channel errors was not considered in this thesis. This is another topic which could be investigated. The optimum quantization strategy for side information, whether all-pole or homomorphic, has not been formulated in an

adaptive transform coding context. Other forms of short term spectrum parameterization are also worthy of study.

In conclusion, future speech digitization developments can not afford to ignore the successes of adaptive transform coding techniques. Despite complexity, transform coding demonstrates that good quality speech is possible at data rates between 8-16 kb/sec. Thus adaptive transform coding is likely to become the speech standard by which others are compared at these rates.

REFERENCES

1. J.L. Flanagan et al,"Speech Coding" IEEE Trans. Commun. COM-27, pp. 710-736, April 1979.
2. J.J. Huang, P.M. Schultheiss,"Block Quantization of Correlated Gaussian Random Variables",IEEE Trans. Commun. Syst. CS-11, pp. 289-296,Sept. 196
3. A. Segall,"Bit Allocation and Encoding for Vector Sources" IEEE Trans. Inform. Theory IT-22, pp.162-168, March 1976
4. R. Zelinski and P. Noll,"Adaptive Transform Coding of Speech Signals" IEEE Trans. Acoust., Speech, Signal Processing, ASSP-25, pp. 299-309, Aug. 1977
5. R. Zelinski and P. Noll,"Adaptive Transform Speech Coding at Low Bit Rates" IEEE Trans. Acoust., Speech, Signal Processing, ASSP-27, pp. 89-95, Feb 1979
6. J. M. Tribolet and R.E. Crochiere,"Frequency Domain Coding of Speech", IEEE Trans. Acoust., Speech, Signal Processing, ASSP-27, pp. 512-530 Oct. 1979
7. R.E. Crochiere and J.M. Tribolet, "Frequency Domain Techniques for Speech Coding" J. Accoust. Soc. Am. Vol-66 pp. 1642-1646, Dec 1979
8. R. Cox and R.E. Crochiere,"Real-Time Simulation of Adaptive Transform Coding" IEEE Trans. Acoust., Speech, Signal Processing, ASSP-2 pp. 147-154
9. A.J. Goldberg, L. Cosell, S. Kwon, L. Bergeron, and R. Cheung. "9600 BPS Speech Optimization Study" GTE Sylvania, Inc., Needham Heights, MA. Communications Systems Div. Sept. 1980

10. J. Max,"Quantizing for Minimal Distortion" IRE Trans. Inform. Theory vol. IT-6, pp. 7-12, March 1960.

11. K. Sam Shanmugam, "Comments on "Discrete Cosine Transform"", IEEE Trans. Comput., Vol C-24, p. 759 July 1975

12. N. Ahmed, T. Natarajan and K. R. Rao,"Discrete Cosine Transform", IEEE Trans. Comput. C-23 pp.90-93 January 1974

13. W. Chien, C. Harrison Smith, and S. C. Fralick "A Fast Computational Algorithm for the Discrete Cosine Transform" IEEE Trans. Commun. COM-25 pp. 1004-1009, Sept. 1977

14. M.J. Narasimha and Allen M. Peterson "On the computation of the Discrete Cosine Transform" IEEE Trans. Commun. COM-26 pp. 934-936 June 1978

15. R.M. Haralick "A Storage Efficient Way to Implement the Discrete Cosine Transform" IEEE Trans. Comput., C-25 pp.764-765, July 1976

16. D. Sloan,"Adaptive Transform Coding of Speech" M. Eng. Thesis Department of Electrical Engineering, McGill University, July 1979

17. R. Viswanathan, J. Makhoul " Quantization Properties of Transmission Parameters in Linear Predictive Systems "IEEE Trans. Acoust., Speech and Signal Process., ASSAP-23, pp 309-321, June 1975