SPEECH TRANSFORM CODING USING RANKED VECTOR QUANTIZATION

by

Ying Zhang

B.A.Sc. Beijing University of Post and Telecommunication, 1983, China
M.A.Sc. Northern Jiao-Tong University, 1988, China

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF APPLIED SCIENCE
in the School
of
Engineering Science

SIMON FRASER UNIVERSITY
July 1996

APPROVAL

Name: Ying Zhang

Degree: Master of Applied Science

Title of Thesis: Speech Transform Coding Using Ranked Vector Quantization

Examining Committee: Dr. Shawn Stapleton
Chairman, Associate Professor

Dr. Vladimir Cuperman
Senior Supervisor, Professor

Dr. Paul Ho
Supervisor, Associate Professor

Dr. Jim Cavers
Internal Examiner, Professor

Date Approved: July 15 1996

PARTIAL COPYRIGHT LICENSE

I hereby grant to Simon Fraser University the right to lend my thesis, project or extended essay (the title of which is shown below) to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its usrs. I further agree that permission for multiple copying of this work for scholarly purposes may be granted by me or the Dean of Graduate Studies. It is understood that copying or publication of this work for financial gain shall not be allowed without my written permission.

Title of Thesis/Project/Extended Essay

"Speech Transform Coding Using Ranked Vector Quantization"

Author: _____
(signature)

(name)

July 15 , 1996
(date)

ABSTRACT

As the demand for mobile communications continues to grow, speech codec designers are faced with the challenge of providing high quality speech coding systems at low bit rate. New efficient speech coding algorithms are required to reduce the bit rates and obtain high quality reproduced speech signal.

Transform coding is a frequency-domain coding technique which has been studied extensively and used widely in low bit rate speech coding systems. The Vector Transform Quantization (VTQ) system is an example of transform coding, where a set of vector quantizers are used to quantize the transform coefficients.

With the motivation of developing high quality speech coders at low bit rate, this thesis investigates two new speech coding algorithms with the goal of obtaining high quality synthetic speech at the rate 2.4 kbps. Based on the VTQ system, the Vector Transform Quantization with Coefficient Ranking (VTQ-CR) system and its enhanced version, Vector Transform Quantization with Coefficient Ranking and Adaptive Linear Prediction (VTQ-CR-ALP), are developed. Coefficient ranking technique and adaptive transform domain linear prediction analysis are proposed to improve the performance of conventional VTQ coders.

The experimental results indicate that ranking transform coefficients in a descending order of their energy values and vector quantizing the most significant coefficients can make the VQ more efficient at low bit rate. A further performance improvement can be achieved by applying an adaptive linear predictor to the voiced ranked coefficients, where the correlations between the coefficients are reduced. Multi-Stage VQ (MSVQ) coupled with the closed-loop VQ codebook search is used to obtain an efficient, high quality and low complexity quantization.

Dedication

I dedicate this thesis to my husband, parents and son for their great love.

Acknowledgements

I would like to thank my supervisor, Dr. Vladimir Cuperman for his support and guidance throughout the course of this thesis project. Thanks to everyone in the speech group for a memorable two year plus.

Contents

List of Tables

List of Figures

List of Abbreviations

ATC	Adaptive Transform Coding
CELP	Code-Excited Linear Prediction
CS-ACELP	Conjugate Structure-Algebraic Code-Excited Linear Prediction
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
FCC	Federal Communication Committee
FFT	Fast Fourier Transform
GLA	Generalized Lloyd Algorithm
ISDN	Integrated Services Digital Network
KLT	Karhunen-Loeve Transform
LD-CELP	Low Delay Code-Excited Linear Prediction
LPC	Linear Prediction Coding
LSP	Line Spectrum Pair
MBE	Multiband Excitation

MELPC	Mixed Excitation Linear Prediction Coding
MPLPC	Multi-Pulse Linear Prediction Coding
MSE	Mean-Squared Error
MSVQ	Multi-Stage Vector Quantization
PCM	Pulse-Code Modulation
SEGSNR	Segmented Signal-to-Noise Ratio
SNR	Signal-to-Noise Ratio
STC	Sinusoidal Transform Coding
TFI	Time-Frequency Interpolation Coding
VQ	Vector Quantization
VTQ	Vector Transform Quantization
VTQ-CR	Vector Transform Quantization with the Coefficient Ranking
VTQ-CR-ALP	Vector Transform Quantization with the Coefficient Ranking

Chapter 1

Introduction

1.1 Background

Speech coding is one of the most crucial parts of advanced digital telecommunication systems. The very high demand for mobile communications is currently the main driving force behind the development of new speech coding algorithms. Offering higher capacities for voice communications at reduced costs, low bit rate speech compression has become an area of intensive research in many telecommunication applications, such as international public telephone networks, digital cellular and mobile communications, video conference, Integrated Services Digital Network (ISDN) and multimedia applications. With the anticipated sharp growth of mobile, personal and secure wireless communications, future wireless digital systems need to rely on advanced source coding technologies for coding speech at rates of 2.4 kbps and below. During the last decade, a great deal of research has been conducted on low bit rate speech coding algorithms and a rapid progress has been made in producing high quality speech at low bit rates [1-10].

The design objective of low bit speech coders is to minimize the bit rate to represent the speech signal while preserving high quality and minimizing the complexity of implementation. Most voice coders (vocoders) operating at rates of around 2.4 kbps are parametric, usually attempting to track the underlying process that generates the speech waveform, by encoding the parameters that describe that process and reproducing the property and resonance of the original speech at the decoder. Linear Predictive Coding (LPC) is one of the most widely used speech coding techniques today [11][12]. LPC analysis may be used to derive the coefficients of a linear digital filter which models the spectral shaping of the vocal track and gives the best spectral match to the speech being encoded. LPC-10 (LPC-10e is its enhanced version) [13][14] and the Mixed Excitation LPC (MELPC) vocoder [15][16] are two examples of LPC vocoder at 2.4 kbps. LPC-10 is a U.S. government standard speech coder for secure communications. The synthesis model is a 10th-order lattice filter, controlled by a set of 10 reflection coefficients which are updated every 22.5ms. The speech produced by the LPC-10 coding algorithm is intelligible, though of a quality significantly lower than telephone quality. The main defect is its perceived robotic quality, which can often disguise the identity of the speaker. The most annoying aspect of the basic LPC vocoder is that, although acceptably intelligible, it does not always sound like natural human speech, especially in the presence of acoustic background noise. MELPC is based on the traditional LPC model of exciting an all-pole filter with either a periodic impulse train to represent voiced speech or white noise to represent unvoiced speech. An improvement on the quality is achieved by using mixed pulse and noise excitation, periodic and aperiodic pulses, a pulse dispersion filter, and

adaptive spectral enhancement [16]. It produces more natural sounding synthetic speech than the LPC-10e vocoder [15].

The introduction of the Multi-Pulse Linear Predictive Coding (MPLPC) in 1982 [1] and the Code-Excited Linear Prediction (CELP) in 1984 [2][3] led to research in a new class of analysis-by-synthesis speech coders [4-6] that enabled us to encode high quality speech at bit rates as low as 4.8 kbps [7][8]. Analysis-by-synthesis, as its name implies, is the use of synthesis as an integral part of the analysis process. The MPLPC is an analysis-by-synthesis system in which each excitation vector consists of a combination of a given number of pulses whose positions and amplitudes are optimized in the closed loop. The derivation of the appropriate pulse positions and amplitudes at the encoder is crucial to the coder performance.

The CELP system uses an 'innovation sequence' as the excitation, instead of using multipulses. At the encoder there is a codebook which contains many different excitation sequences (innovations). The 'optimum' innovation sequence is chosen to minimize a given distortion criterion between the original and the synthesized speech. At the decoder of a CELP codec, each block of reconstructed speech samples is produced by filtering the selected innovation sequence through a long-term filter and then a LPC vocal tract filter. CELP encoding is a very important technique--several speech coding standards, such as the 4.8 kbps US Department of Defense standard FS-1016 for secure telephony [17], the ITU G.728 LD-CELP [18] 16 kbps standard, and the ITU G.729 CS-ACELP [19] 8 kbps standard, are all based on the CELP coders. Although both multi-pulse and CELP excitation models use analysis-by-synthesis techniques and represent very important steps

in synthesizing natural-sounding speech, these models are not able to reduce the bit rate for high quality speech below 4 kbps. Due to the lack of the efficient presentation of the excitation, the performance of CELP coders degrades rapidly below 4.8 kbps [20].

Transform coding is a frequency-domain coding technique which has been studied extensively and used widely [21]. One of the well-known types of transformations is the orthogonal transform. There are two reasons for using orthogonal transform. The first is that such a transform can help to reduce the bit rate since it distributes the signal power non-uniformly over the transform coefficients; the second reason is that such a transform can decompose the signal into perceptually relevant components. By quantizing these components with different accuracies, the coding errors can be controlled in such a way that perceptible distortion is minimal [21][22]. The bit rate required to encode the transformed samples is lower than the rate required to encode the untransformed samples to obtain the same quality synthesized speech.

If the input signal samples are Gaussian distributed random variables and the transform coefficients are scalar quantized, the Karhunen-Loeve Transform (KLT) was proven to be optimal orthogonal transform under the minimum mean-squared error (MSE) criterion [21][23][56]. The KLT completely decorrelates the signal sequence in the transform domain. However, KLT is signal-dependent and difficult to implement. The other two well-known orthogonal transforms are the Discrete Cosine Transform (DCT) which is used in the Adaptive Transform Coding (ATC) [23] and the Discrete Fourier Transform (DFT) which is used in the Multiband Excitation (MBE) coding model [24]. Both DCT and DFT are signal-independent and sub-optimal in the sense that they cannot fully

decorrelate the transform coefficients. In the application in speech coding, DCT demonstrates a better performance than the DFT [23][21]. Experimental results indicate that DCT is the closest in performance to KLT [21]. Moreover, there are fast computational algorithms for DCT that allow easy VLSI implementation.

In ATC, although the total number of bits available to quantize the transform coefficients remains constant, the bit allocation to each coefficient changes from frame to frame. This dynamic bit allocation is controlled by the time-varying statistics of the speech, which have to be transmitted as the side information. The number of bits assigned to each transform coefficient is depends on its corresponding spectral energy value. ATC was used to encode speech successfully at bit rate in the range of 9.6 - 20 kpbs.

MBE model assumes that both voiced and unvoiced excitation can exist at the same time in the same analysis frame but in different frequency bands. The speech spectrum is split into non-overlapping bands and each band is modeled as being either voiced or unvoiced. The voiced bands are synthesized using sinusoidal signal and the unvoiced bands are synthesized using bandpass filtered noise. The required frequency band analysis is obtained by using a DFT.

Other important speech coding algorithms for 2.4 kbps include the Time-Frequency Interpolation Coding (TFI) [25] and the Sinusoidal Transform Coding (STC) [26], etc. The STC algorithm uses a sinusoidal model with amplitudes, frequencies, and phases derived from a high resolution analysis of the short-term Fourier transform of the speech signal. A harmonic set of frequencies is used to represent the periodicity of the input speech. Pitch, voicing, and sine wave amplitudes are transmitted to the decoder.

Conventional methods are used to code the pitch and voicing, and the sine wave amplitudes are coded by fitting a set of cepstral coefficients to an envelope of the amplitudes.

Many transform coding systems use scalar quantization [23], where the transform coefficients or speech analysis parameters (linear prediction coefficients etc.) are quantized individually by a set of scalar quantizers. It was proven that the vector quantizer can achieve better performance than the scalar quantizer [27]. The idea of vector quantization (VQ) is to map each input vector into a codevector in a codebook which consists of a finite set of vectors and covers the anticipated range of the input values of the quantizer. In each analysis interval, the codebook is searched, the codevector which gives the best match to the input is selected and the corresponding index is transmitted. The interest in VQ resides in the fact that additional gain in performance is achievable over scalar quantization even in the case the input vector consists of independent elements. One of greatest difficulties with VQ is in setting up a good quality codebook; significant amounts of training are involved. In addition, there is a need for very efficient algorithms to keep the codebook search complexity reasonable. In the last decade, many efficient algorithms for searching VQ codebooks have been developed [28-30].

As the demand for mobile communications continues to grow, speech codec designers are faced with great challenges in providing high quality speech coding systems at low bit rate and low costs. At the rate 2.4 kbps, the number of bits available to produce a signal that matches the original speech waveform is insufficient. An efficient and effective speech coding algorithm is required to reduce information loss and obtain high quality reproduced

speech signal. Because of the model inadequacies in current 2.4 kbps vocoders, the quality of reconstructed speech is still lower than desired. With the motivation of developing high quality speech coders at low bit rate, this thesis is concerned with investigating two new speech coding algorithms with the goal of obtaining high quality synthetic speech at the rate 2.4 kbps.

A typical Vector Transform Quantization (VTQ) system is a coding system where each consecutive M samples of a waveform are transformed into a set of coefficients which are quantized by a set of $m<<M$ vector quantizers. Based on the VTQ system, the Vector Transform Quantization with the Coefficient Ranking (VTQ-CR) system and its enhanced version, Vector Transform Quantization with the Coefficient Ranking and Adaptive Linear Prediction (VTQ-CR-ALP) system, are introduced in this thesis. In these two systems, before the vector quantization, the transform coefficients are ranked in a descending order of their energy values. Only the first N_t ranked transform coefficients with higher energy values are quantized. The order information is extracted from the short-term spectral information of the speech. The voiced and unvoiced frames are encoded by searching different codebooks. The decoder restores the order of the quantized coefficient sequence and takes an inverse transform to reconstruct the corresponding block of samples. Moreover, in the VTQ-CR-ALP system, an adaptive transform domain linear predictor is applied to the voiced ranked transform coefficients, improving the system performance further by reducing their near-sample correlations. The ranked transform coefficients are reconstructed at the decoder from the knowledge of the short-term speech spectrum. The analysis-by-synthesis method is used to encode the excitation of the synthesizer. The

performances of these two systems are evaluated along with a conventional VTQ system. The experimental results indicate that the transform coefficient ranking and the adaptive transform domain linear prediction are efficient methods in improving the performance of the transform coding system. System simulation results are presented for a Gauss-Markov source and speech source.

1.2 Contributions of the Thesis

The major contributions of this thesis can be summarized as follows:

1. The development of new speech coding algorithms at the bit rate 2.4 kbps; the features of the coding schemes include the use of the order ranking vector quantization to encode the transform coefficients and the application of the adaptive transform domain linear prediction in the transform coding system.

2. The design and the analysis of the proposed coder with the optimal bit allocation scheme.

3. Simulations of these coding systems. The performances of all these systems are compared and evaluated. It is shown that the new systems improve the performance, when compared to the known VTQ approach.

1.3 Thesis Outline

The rest of this thesis is organized as follows. In Chapter 2, the basic techniques used in speech coding are reviewed. A vector transform quantization (VTQ) system at 2.4 kbps is described in detail in Chapter 3. The optimal bit assignment strategy is discussed. Two new coding systems with the coefficient ranking strategy, the VTQ-CR and the VTQ-CR-ALP, are presented in Chapter 4. The adaptive transform domain linear prediction is applied in the VTQ-CR-ALP system and the analysis-by-synthesis method is used to quantize the excitation of the synthesizer. In Chapter 5, simulation results of all speech coders in this thesis are presented and the performances are evaluated. Finally, conclusions are drawn in Chapter 6.

Chapter 2

Analysis and Compression of Digital Speech

The purpose of this chapter is to present an overview of the digital signal processing techniques related to the analysis and compression of the speech signal. From Section 2.1 to section 2.7, some important concepts for the research presented in this thesis are reviewed, including the characteristics of the voiced and unvoiced speech, the scalar quantization and vector quantization, the analysis-by-synthesis technique and the linear prediction analysis, and transform representation of speech etc. In Section 2.8, two classes of coding systems, analysis-by-synthesis speech coding and transform coding are discussed.

2.1 Voiced Speech and Unvoiced Speech

For the voiced speech, the signal waveform is considered periodic at a rate corresponding to the glottal pulse frequency. The period may be variable over the duration of a speech segment, and the shape of the periodic wave usually changes gradually from segment to segment. For the unvoiced speech, the signal is like random noise, being produced by the turbulent flow of air at restrictions in the vocal tract. However, the spectrum of unvoiced speech does not have a truly flat energy spectrum as would Gaussian white noise; the spectrum is shaped by the resonance of the vocal tract, and this gives rise to a small amount of predictability.

Fig 2.1 and Fig 2.2 are examples of voiced speech and unvoiced speech, respectively.

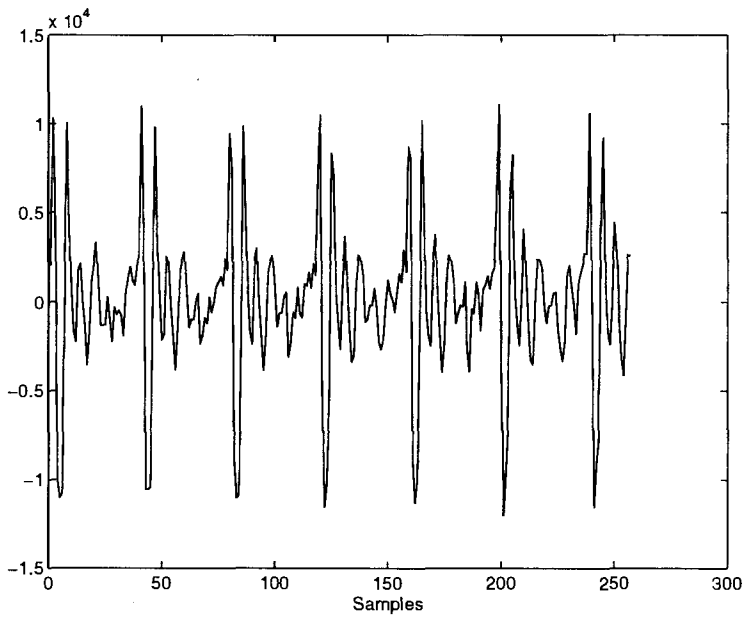


Fig 2.1 Voiced Speech

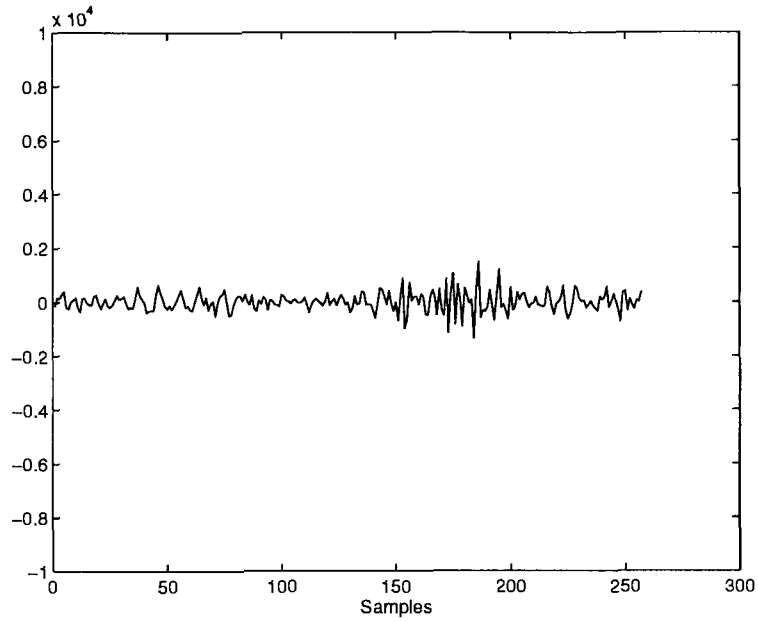


Fig 2.2 Unvoiced Speech

In the speech signal analysis, we assume that the speech waveform can be considered stationary over a sufficiently short time interval, although it is non-stationary over a long interval of time. In order to handle the time-variant characteristic of the speech signal, the spectral information of the speech signal is transmitted every 20-30ms.

2.2 Scalar Quantization and Vector Quantization

Vector quantization (VQ) can be viewed as a mapping of a vector in a k-dimensional Euclidean space, $\mathbf{R}^k$, into a codevector in a finite set, C, containing N codevectors such that the difference between the input vector and the quantized vector is minimized according to some chosen criteria [28]. The set C is called a codebook.

$$C = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N\} \qquad \mathbf{c}_i \in \mathbf{R}^k \tag{2.1}$$

The most common distortion measure is the squared Euclidean distance

$$d(\mathbf{y}, q(\mathbf{y})) = \|\mathbf{y} - q(\mathbf{y})\|^2 \tag{2.2}$$

where $q(\cdot)$ is the quantization operator and $\mathbf{y}$ is the input vector. A vector quantizer defines a partition of N cells in the k-dimensional space which are denoted by Ω_i, $i = 1,2,\ldots,N$, and associates each cell Ω_i with a codevector $\mathbf{c}_i$. The partition satisfies

$$\Omega_i \cap \Omega_j = \varnothing \quad \text{for } i \neq j; \tag{2.3}$$

and

$$\bigcup_{i=1}^{N} \Omega_i = \mathbf{R}^k \tag{2.4}$$

The quantizer assigns the codevector $\mathbf{c}_i$ if the input $\mathbf{y}$ belongs to the Ω_i. There are two necessary conditions for an optimal quantizer. The first is the *nearest neighbor condition*

$$q(\mathbf{y}) = \mathbf{c}_i, \quad \text{iff} \quad d(\mathbf{y}, \mathbf{c}_i) \leq d(\mathbf{y}, \mathbf{c}_j), \quad j \neq i, \quad 1 \leq j \leq N \tag{2.5}$$

which means that a VQ chooses a codevector that results in the minimum distortion with respect to the input vector $\mathbf{y}$. The second condition for optimality is that each codevector $\mathbf{c}_i$ is chosen to minimize the average distortion in the cell Ω_i. For the squared Euclidean distance, such a vector is found to be the centroid of the cell Ω_i, i.e.,

$$\mathbf{c}_i = E[\mathbf{y} | \mathbf{y} \in \Omega_i] \tag{2.6}$$

This condition is called the *centroid condition.* If a set of input vectors called the training set is given, an iterative training procedure known as the generalized Lloyd algorithm (GLA) can be used to design a VQ codebook [28].

In speech coding systems employing VQ, identical codebooks are stored in both the encoder and the decoder. For each input vector $\mathbf{y}$, the quantization is performed by computing the distortion between $\mathbf{y}$ and each of the codevectors, then choosing the codevector with the minimum distortion as the quantized value of $\mathbf{y}$. This type of VQ is known as the full search since all codevectors are tested for quantizing each input vector. The index of the selected codevector is transmitted to the decoder where it is used as an entry to obtain the corresponding quantized vector from the decoder's copy of the codebook.

Vector quantization can offer substantial performance advantages over scalar quantization at very low bit rates. However, these advantages are obtained at considerable computational and storage costs. For the full search quantization, the complexity increases exponentially as a function of the number of bits per vector. It is the computational complexity and storage requirement of a VQ codebook that have led speech researchers to develop a number of structured VQ schemes [28-30] in which structure is added to the codebook in the design process to reduce the computational and/or memory complexity, such as the multi-stage VQ, the gain-shape VQ, and the tree-searched VQ.

A structured VQ scheme which can achieve very low encoding and storage complexity in comparison to unstructured VQ is the multistage vector quantization (MSVQ) [31][28]. The basic idea of the MSVQ is first to perform a relatively crude quantization by using a

small codebook and then to provide a further refinement by using successive codebooks. The available B_T code bits are divided among L stages, with B_i bits for stage i, $i \in \{1,\ldots,L\}$, and $\sum_{i=1}^{L} B_i = B_T$. Each stage consists of a codebook with 2^{B_i} codevectors. A representation $\hat{\mathbf{y}}$ of an input vector $\mathbf{y}$ is formed by selecting from each stage codebook a codevector, $\hat{\mathbf{y}}_i$ for stage i, and forming their sum, i.e., $\hat{\mathbf{y}} = \sum_{i=1}^{L} \hat{\mathbf{y}}_i$. Thus, the storage complexity of an MSVQ is $\sum_{i=1}^{L} 2^{B_i}$ vectors, which can be much less than the complexity of $\prod_{i=1}^{L} 2^{B_i} = 2^{B_T}$ vectors for an unstructured VQ. In a traditional MSVQ, the stage codebooks are searched sequentially. In each stage, a residual vector is generated and passed to the next stage to be quantized independently of the other stages. The encoding complexity for sequential searching of the stage codebooks is $\sum_{i=1}^{L} 2^{B_i}$. In spite of the complexity advantages of the MSVQ, the conventional stage-by-stage design of the codebooks in the MSVQ is suboptimal with respect to the overall performance measure. The performance of the conventional multistage vector quantizer is usually lower than that for a single-stage vector quantizer with the same number of bits and it can be enhanced by joint design of the MSVQ codebooks and a tree-search strategy[32][33].

There are two other methods for reducing the VQ complexity: split VQ and gain-shape VQ. In a split VQ, the input vector is partitioned and quantized by separate VQs. In a gain-shape VQ, the input vector is normalized by a gain and the codebook contains

vectors for which only the shape varies. Thus for input vectors with different amplitudes but the same shapes, only one codevector has to be available. This substantially reduces the number of required codebook vectors.

Scalar quantization is the independent quantization of each signal value or parameter. It can be viewed as a special VQ with dimension equal to one.

2.3 Analysis -by- Synthesis

Analysis-by-Synthesis is a general approach for estimating a set of parameters of a speech production model. The model is assumed to be able to generate a variety of speech waveforms by adjusting the parameters; the synthesized speech signals are compared to the original speech signal, and the model parameters are varied in a systematic way to obtain the best match between the original and the synthesized signal. The first application of the analysis-by-synthesis technique to speech coding is due to Atal and Remde [1].

In the case of applying the analysis-by-synthesis technique to low rate speech coding, the speech production model is used at the encoder to find the optimal set of parameters for reproducing each segment of the original speech signal under a given distortion criterion. Then, the optimal parameters are transmitted to the decoder which uses an identical speech production model and the received set of parameters to synthesize the speech waveform. In other words, the incorporation of a decoder within the speech encoder allows the selection of the set of parameters which minimizes in some sense the error

between the original speech and the synthesized speech. Coding the parameters, rather than the entire speech waveform can result in a significant data compression ratio. At the same time, the fact that parameters choice at the transmitter is based on direct comparison of the reconstructed and original waveforms helps to preserve good speech quality at low rates.

Fig 2.3 is a block diagram of an analysis-by-synthesis system.

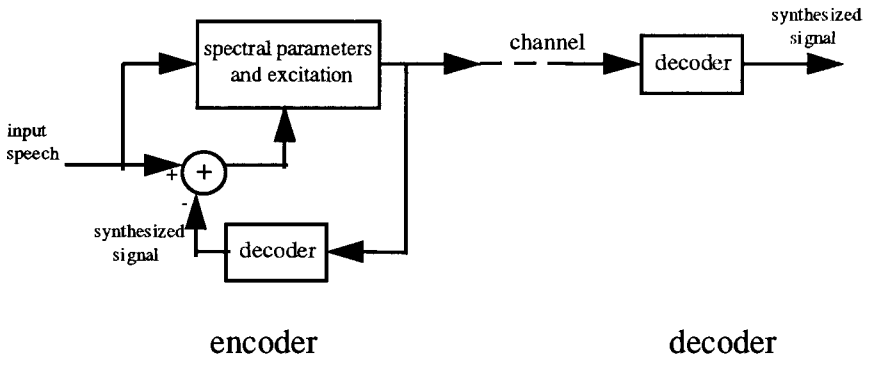


Fig 2.3 Block diagram of an analysis-by-synthesis system

2.4 Linear Prediction

Linear prediction is a very important and powerful speech processing technique. The basic idea behind the method is that sample values of speech, x_n, can be approximated as a linear combination of the past p speech samples. Mathematically, the linear predictor is described by

$$\hat{x}_n = \sum_{i=1}^{p} a_i x_{n-i} \tag{2.7}$$

where $\hat{x}_n$ is the predicted sample at instant n. a_i's, $i = 1,2,\ldots,p$, are the linear prediction coefficients and p is the predictor order. These coefficients are determined by minimizing the mean-squared error (MSE) between the actual speech samples and the linearly predicted ones.

$$\sigma_e^2 = E\left[e_n^2\right] = E\left[(x_n - \hat{x}_n)^2\right] \tag{2.8}$$

Setting the partial derivative of σ_e^2 with respect to each coefficients a_i, $i = 1,2,\ldots,p$ to zero results in a set of p linear equations with p unknowns a_i, which can be written as

$$r_{xx}(n) = \sum_{j=1}^{p} a_j r_{xx}(|n-j|) \qquad n = 1,2,\ldots,p \tag{2.9}$$

where $r_{xx}(n)$, $n = 0,1,2,\ldots,p$, is the autocorrelation function of the input signal. In a matrix form, the equations in (2.9) becomes

$$\mathbf{R}_{xx}\mathbf{a} = \mathbf{r}_{xx} \tag{2.10}$$

where $\mathbf{R}_{xx}$ is the autocorrelation matrix,

$$\mathbf{R}_{xx} = \begin{bmatrix} r_{xx}(0) & r_{xx}(1) & r_{xx}(2) & \cdots & r_{xx}(p-1) \\ r_{xx}(1) & r_{xx}(0) & r_{xx}(1) & \cdots & r_{xx}(p-2) \\ r_{xx}(2) & r_{xx}(1) & r_{xx}(0) & \cdots & r_{xx}(p-3) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{xx}(p-1) & r_{xx}(p-2) & r_{xx}(p-3) & \cdots & r_{xx}(0) \end{bmatrix} \tag{2.11}$$

and $\mathbf{a} = [a_1, a_2, \ldots, a_p]^T$, $\mathbf{r}_{xx} = [r_{xx}(1), r_{xx}(2), \ldots, r_{xx}(p)]^T$. Hence, the solution of (2.10) is given by

$$\mathbf{a} = \mathbf{R}_{xx}^{-1}\mathbf{r}_{xx} \tag{2.12}$$

Equation (2.10) is called the Yule-Walker equation [28]. Fig 2.4 shows the model of a linear predictor. With input x_n and output e_n, the z-domain transfer function is

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i} \tag{2.13}$$

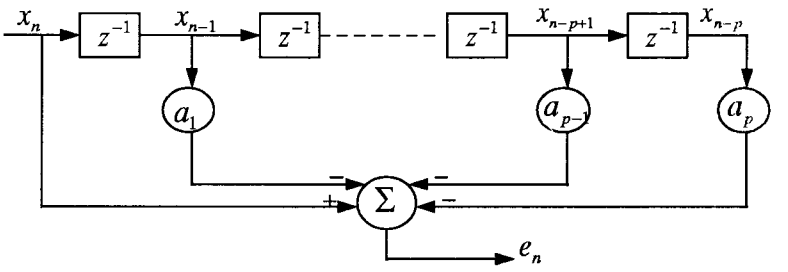


Fig 2.4 The model of a linear predictor

It can be shown that for a stationary random process, the prediction error of the optimal infinite-order linear predictor is a white noise process. $A(z)$ is commonly referred to as a whitening filter. The inverse filter $1/A(z)$ can convert e_n back to the original signal x_n.

Since the local stationary model of the speech signal is considered in speech analysis, a set of predictor coefficients is computed every 20-30 ms in order to match the time-varying properties of the speech signal. A value of prediction order p ranging from 10 to 20 is normally sufficient for a good spectral estimate of a speech signal. The

autocorrelation method can be used to find the short-time autocorrelation values [34]. For a segment of speech signal, $x_0, x_1, \ldots, x_{L-1}$, the autocorrelation functions are estimated by

$$\tilde{r}_{xx}(n) = \frac{1}{L} \sum_{i=0}^{L-n-1} x_i x_{n+i} \qquad n = 0,1,2,\ldots,L-1 \qquad (2.14)$$

The successive segments of speech signal are windowed by multiplying the signal with a rectangular window which has a value of 1 in the interval *(0, L-1)* and is 0 outside. Typically, a smooth window, w_i, such as the Hamming window [28] is used in order to obtain a better spectral estimate. In this case, the autocorrelation functions in (2.14) become

$$\tilde{r}_{xx}(n) = \frac{1}{L} \sum_{i=0}^{L-n-1} x_i w_i x_{n+i} w_{n+i} \qquad n = 0,1,2,\ldots,L-1 \qquad (2.15)$$

The autocorrelation matrix is Toeplitz and symmetrical. The Levinson-Durbin algorithm [35-37] can be used for finding the linear predictive coefficients. This method requires much less computational effort than a general method for solving the set of linear equations.

2.5 Line Spectrum Pairs (LSP)

Direct quantization of the linear prediction coefficients is known not to be efficient and can lead to an unstable inverse filter $1/A(z)$. We therefore transform these coefficients to an equivalent set of parameters, such as the reflection coefficients or the line spectrum

pairs (LSPs). Among the various linear prediction coefficient representations, the LSPs are the most efficient parameters for quantization while maintaining stability. Using this technique, the transfer function of the analysis filter is represented by two functions which have their zeros on the unit circle.

For the transfer function of the LPC analysis filter in (2.13), we define

$$P(z) = A(z) - z^{-(p+1)}A(z^{-1}) \tag{2.16}$$

$$Q(z) = A(z) + z^{-(p+1)}A(z^{-1}) \tag{2.17}$$

where p is the order of the linear predictor. $P(z)$ has zeros $z = 1$ and $z_k = e^{j2\pi f_k T_s}$, $k = 1,2,\ldots,{}^{p}/_{2}$. $Q(z)$ has zeros $z = -1$ and $z_k = e^{j2\pi g_k T_s}$, $k = 1,2,\ldots,{}^{p}/_{2}$ [38][39]. f_k and g_k are frequencies of zeros, T_s is the sample time. With $A(z)$ known, f_k's and g_k's can be obtained. f_k and g_k make up the kth line spectral pair. If the inverse filter ${}^{1}/_{A(z)}$ is stable, LSPs alternate on the frequency scale [38][39], i.e.

$$f_1 < g_1 < f_2 < g_2 \ldots < f_{p/2} < g_{p/2} \tag{2.18}$$

This property can be used to check the stability of the inverse filter ${}^{1}/_{A(z)}$.

After vector quantization, the LSPs are transmitted to the receiver. At the receiver, they are converted back to the linear prediction coefficients by using the following equations:

$$P(z) = (1 - z^{-1})\prod_{i=1}^{p/2}(1 - 2z^{-1}\cos(g_i) + z^{-2}) \tag{2.19}$$

$$Q(z) = (1+z^{-1})\prod_{i=1}^{p/2}(1-2z^{-1}\cos(f_i)+z^{-2}) \tag{2.20}$$

$$A(z) = \frac{1}{2}[P(z)+Q(z)] \tag{2.21}$$

2.6 Pitch extraction

Pitch estimation is an essential requirement in speech coding systems. Many different algorithms have been investigated and most of them work well on strongly voiced sounds. Many of the algorithms were discussed by Rabiner [40]. In the case of voiced speech, the short-time autocorrelation function exhibits peaks at time-shifts corresponding to multiples of the pitch-period. At these points the speech signal is in phase with the delayed version of itself, giving high correlation values. This suggests that the short-time autocorrelation function is a powerful technique for estimating the pitch period of voiced speech. The short-time autocorrelation function forms the basis of some pitch extraction algorithms [40], in spite of the fact that there are a number of efficient time-domain algorithms which operate directly on the time-waveform.

2.7 Transform Representation of Speech

Orthogonal transform plays an important role in the analysis and compression of the speech signal. Orthogonal transforms are block-based operations. The input signal is

segmented in blocks of M samples. Each block is treated as a vector $\mathbf{x}$, $\mathbf{x}=[x_0, x_1, \ldots, x_{M-1}]^T$ which is transformed to another representation, denoted by vector $\mathbf{s}$, $\mathbf{s}=[s_0, s_1, \ldots, s_{M-1}]^T$,

$$\mathbf{s} = \mathbf{T}\mathbf{x} \tag{2.22}$$

The components of $\mathbf{s}$ are usually called the transform coefficients. The transform matrix $\mathbf{T}$ is orthogonal, therefore, its inverse is given by

$$\mathbf{T}^{-1} = \mathbf{T}^T \tag{2.23}$$

The superscript T denotes the matrix transposition.

If the input signal samples are Gaussian distributed random variables and the transform coefficients are scalar quantized, the Karhunen-Loeve Transform (KLT) was proven to be optimal orthogonal transform under the minimum mean-squared error (MSE) criterion [21][23][56]. The orthogonal basis functions for KLT are obtained as the eigenvectors of the covariance matrix of the input signal.

Let $\mathbf{x}$ be a zero mean M-dimensional random input vector, $\mathbf{x}=[x_0, x_1, \ldots, x_{M-1}]^T$, the transform vector $\mathbf{s}$ is given by

$$\mathbf{s}=\mathbf{U}\mathbf{x} \tag{2.24}$$

where $\mathbf{U}$ is the KLT matrix, $\mathbf{U}=[\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{M-1}]^T$. $\mathbf{u}_i$'s, $i=0,1,\ldots,M-1$, are eigenvectors of the autocorrelation matrix of $\mathbf{x}$. They are orthogonal and satisfy

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

Since the autocorrelation matrix of **x** is symmetric and positive, it can be decomposed as

$$\mathbf{R}_{xx} = E[\mathbf{x}\mathbf{x}^T] = \mathbf{U}^T \Lambda \mathbf{U} \tag{2.25}$$

where $\mathbf{U}^T$ is the eigenvector matrix of $\mathbf{R}_{xx}$ and Λ is the eigenvalue diagonal matrix of $\mathbf{R}_{xx}$. Therefore, the autocorrelation matrix of the transform vector **s** is equal to

$$\begin{aligned} \mathbf{R}_{ss} &= E[\mathbf{s}\mathbf{s}^T] = E[\mathbf{U}\mathbf{x}\mathbf{x}^T\mathbf{U}^T] \\ &= \mathbf{U}\mathbf{R}_{xx}\mathbf{U}^T \\ &= \Lambda \end{aligned} \tag{2.26}$$

Here, $\mathbf{U}\mathbf{U}^T = \mathbf{I}_M$, $\mathbf{I}_M$ is a $M \times M$ identity matrix. Therefore, the KLT completely decorrelates the signal in the transform domain.

Although KLT gives the best MSE performance, its main disadvantage is the lack of fast algorithms to compute the transform coefficients. The basis functions of KLT are dependent on the autocorrelation matrix of the input signal and a lot of computation is needed to determine the eigenvectors. This has made KLT an ideal transform but impractical tool. However, the KLT does provide a benchmark against which other discrete transforms may be judged.

Another important orthogonal transform is the Discrete Cosine Transform (DCT) which is defined as [21]:

$$\mathbf{s} = \mathbf{A}\mathbf{x} \tag{2.27}$$

where **A** is an $M \times M$ transform matrix whose k th row, n th column element a_{kn} is given by

$$a_{kn} = \alpha(k)\cos\left[(2n+1)\frac{\pi k}{2M}\right], \qquad k,n = 0,1,\ldots,M-1 \qquad (2.28)$$

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

DCT is signal-independent and a suboptimal orthogonal transform. A very low correlation between transform coefficients is achieved by the DCT. There are other important orthogonal transforms used in the speech signal analysis and compression, such as Discrete Fourier Transform (DFT), Walsh-Hadamard Transform (WHT), etc.. Zelinski and Noll's research results showed that DCT is the closest in performance to KLT among DCT, DFT and WHT [41][21]. Moreover, there are fast computational algorithms for DCT which allow easy VLSI implementation. The DCT is the most popular transform used in the digital speech and image processing [42].

2.8 Speech Coding Systems

Speech coding can generally be classified into two categories: waveform coders and source coders or vocoders. The aim of waveform coding is to reproduce the original signal as accurately as possible. It usually requires a high bit rate (> 16 kbps). Waveform coders are generally signal independent. In contrast, the vocoders extract perceptually

significant parameters from the input signal in order to synthesize a reproduced signal which is acceptable to a human ears. They are based on a model of speech production and hence are signal dependent. Vocoders achieve a higher data compression ratio than waveform coders. Since vocoders make no attempt to reproduce the original waveform, they can operate at very low bit rates ($\leq$ 2.4 kbps) but the lower rates are obtained at the expense of reduced speech quality.

2.8.1 Analysis-by-Synthesis Speech Coding

At low bit rates, the most successful linear predictive based speech coding algorithms are based on analysis-by-synthesis techniques. The Code Excited Linear Predictive Coding (CELP) [2][3] is an example of such coding techniques.

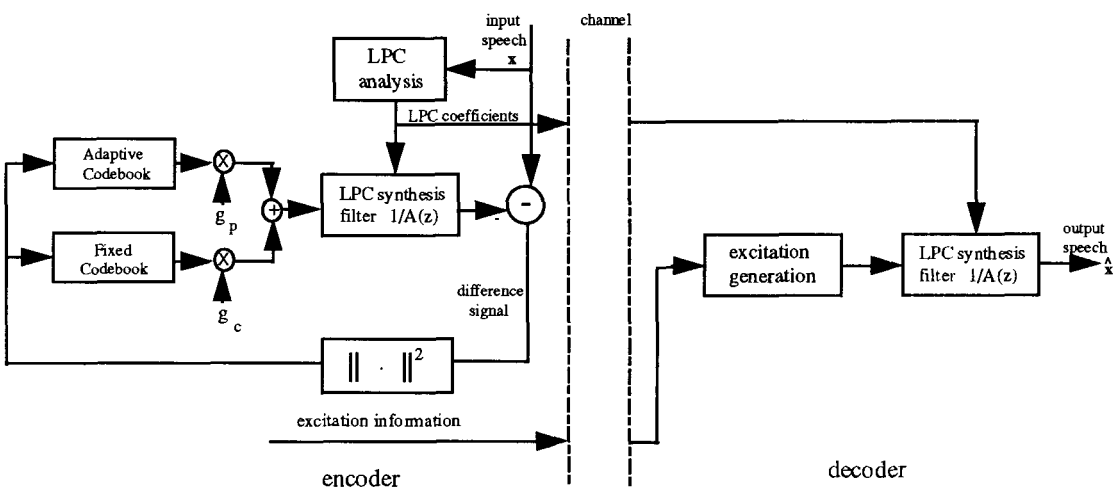


Fig 2.5 An analysis-by-synthesis LPC coding scheme

Fig 2.5 shows a generalized analysis-by-synthesis LPC coding scheme. For an analysis-by-synthesis LPC-based coder, the LPC synthesizer model is employed at both the

encoder and decoder. At the encoder the aim of linear prediction analysis is to extract a set of parameters from the speech signal to specify the synthesizer transfer function which gives the best match to the speech to be coded. These optimum parameters, i.e., predictor coefficients, are obtained by applying the method which is described in Section 2.4. The LPC coefficients are converted to the LSPs for VQ. The transfer function of the LPC analysis filter is given by

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}$$

An all-pole filter $1/A(z)$ is used to model the spectral shaping of the vocal tract. The synthesized speech signal, $\hat{\mathbf{x}}$, is produced by feeding this filter with an excitation signal [12][50]. The current excitation vector should be selected such that the distortion $\varepsilon_{\mathbf{x}}$ between $\mathbf{x}$ and $\hat{\mathbf{x}}$ is minimized.

$$\varepsilon_{\mathbf{x}} = \|\mathbf{x} - \hat{\mathbf{x}}\|^2 \qquad (2.29)$$

In the CELP codec, two different kinds of excitation codebooks are used. One is the adaptive codebook which initiates the quasi-periodicity of the voiced speech, the other is the fixed codebook which provides the innovation excitation. The optimal codevector for the fixed codebook search is determined by minimizing the MSE, ε,

$$\varepsilon = \|\mathbf{t} - \mathbf{y}_i\|^2 \qquad (2.30)$$

where $\mathbf{t}$ is the target vector formed by subtracting the zero-input response (ZIR) of the synthesized filter $1/A(z)$ from the input speech, and $\mathbf{y}_i$ is the zero-state response (ZSR)

of the filter generated using the i th codevector $\mathbf{c}_i$. Let $\mathbf{H}$ be a lower triangle impulse response matrix,

$$\mathbf{H} = \begin{bmatrix} h(0) & 0 & \cdot & \cdot & \cdot \\ h(1) & h(0) & \cdot & \cdot & \cdot \\ h(2) & h(1) & h(0) & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ h(n-1) & h(n-2) & \cdot & \cdot & h(0) \end{bmatrix}$$

then, Equation (2.30) can be written as

$$\varepsilon = \left\| \mathbf{t} - g_c \mathbf{H}\mathbf{c}_i \right\|^2 \tag{2.31}$$

where g_c denotes the gain. Minimizing ε with respect to g_c in Equation (2.31), the optimal gain is found to be

$$g_c = \frac{\mathbf{t}^T \mathbf{H}\mathbf{c}_i}{\left\| \mathbf{H}\mathbf{c}_i \right\|^2} \tag{2.32}$$

If g_c is substituted into (2.31), minimizing the ε is equivalent to maximize ε^*

$$\varepsilon^* = \frac{\left(\mathbf{t}^T \mathbf{H}\mathbf{c}_i\right)^2}{\left\| \mathbf{H}\mathbf{c}_i \right\|^2} \tag{2.33}$$

The index of the selected codevector is transmitted to the decoder in order to construct the excitation vector.

For the adaptive codebook search, the excitation vector is formed by delaying the previous excitation vectors. The optimal delay is the one which generates the excitation vector to maximize ε^* in Equation (2.33).

The sequential codebook search is usually used in the CELP coding system. The adaptive codebook is first searched, then the contribution of the adaptive codebook is subtracted from the target vector before searching the fixed codebook. A lot of work has been done to reduce the codebook search complexity. Most of the fixed codebooks are structured codebooks, instead of the stochastic codebooks, such as the multi-pulse codebook [3] and the algebraic codebook [19].

2.8.2 Transform Coding

Transform coding is a frequency-domain coding technique. In transform coding systems, each block of speech samples is transformed into a set of transform coefficients; these coefficients are then quantized and transmitted. An inverse process in the receiver converts the frequency-domain encoded signal back into the time-domain to obtain the corresponding block of reconstructed speech samples. Fig 2.6 shows the general diagram of a transform coding system.

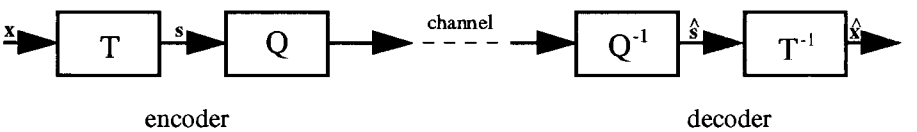


Fig 2.6 Transform coding scheme

At the encoder, the input signal vector **x** is transformed to the vector **s** by an orthogonal transform matrix **T**:

$$\mathbf{s} = \mathbf{T}\mathbf{x}$$

The transform coefficients, which are the elements of $\mathbf{s}$, are quantized and transmitted across the channel. The decoder takes the inverse quantization and transform, then generates a synthesized signal $\hat{\mathbf{x}}$.

$$\hat{\mathbf{x}} = \mathbf{T}^{-1}\hat{\mathbf{s}} \tag{2.34}$$

The quantization error in $\mathbf{s}$, $\mathbf{s}-\hat{\mathbf{s}}$, can results in an error $\mathbf{e}_x$ in the coding system

$$\mathbf{e}_x = \mathbf{x} - \hat{\mathbf{x}} = \mathbf{T}^{-1}(\mathbf{s} - \hat{\mathbf{s}}) \tag{2.35}$$

Let ε be the mean-squared error (MSE) between the original signal $\mathbf{x}$ and the reconstructed one $\hat{\mathbf{x}}$, then

$$\begin{aligned} \varepsilon &= E[(\mathbf{x}-\hat{\mathbf{x}})^T(\mathbf{x}-\hat{\mathbf{x}})] \\ &= E[(\mathbf{s}-\hat{\mathbf{s}})^T\mathbf{T}\mathbf{T}^{-1}(\mathbf{s}-\hat{\mathbf{s}})] \\ &= E[(\mathbf{s}-\hat{\mathbf{s}})^T(\mathbf{s}-\hat{\mathbf{s}})] \end{aligned} \tag{2.36}$$

which means that ε is equal to the mean-squared quantization error of the transform vector $\mathbf{s}$. In other words, the MSE is unaffected by an orthogonal transform.

The orthogonal transform distributes the signal power non-uniformly over the transform coefficients. Therefore, the transform coefficients can be efficiently coded by assigning more bits to the components with higher energy.

Transform coding gain, G_{TC}, is introduced to evaluate the performance of the transform coding. With the scalar quantization, it is defined as the ratio of the distortion of the Pulse-Code Modulation (PCM) coding, D_{PCM}, over the distortion of the transform coding, D_{TC}.

$$G_{TC} = \frac{D_{PCM}}{D_{TC}} \tag{2.37}$$

For the Gaussian source, the optimum bit assignment for the scalar quantization of the transform coefficients in terms of the MSE criterion was derived in [23], which is given by

$$b_i = r + \frac{1}{2}\log_2 \frac{\sigma_i^2}{\left(\prod_{j=1}^{M}\sigma_j^2\right)^{1/M}} \tag{2.38}$$

where b_i is the number of bit assigned to the i th transform coefficient, r is the average bit rate in bits/sample and σ_i^2 is the variance of the i th transform coefficient. If the optimal bit assignment in (2.38) is used for the quantization of the transform coefficients, then

$$D_{TC} = 2^{2\delta}2^{-2r}\left(\prod_{j=1}^{M}\sigma_j^2\right)^{1/M} \tag{2.39}$$

where δ is an item related to the practical quantizer. For the same source, if the PCM is used, the distortion is equal to

$$D_{PCM} = 2^{2\delta}2^{-2r}\sigma^2 \tag{2.40}$$

where, for the orthogonal transform, the variance σ^2 is equal to the average of the variances of the transform coefficients

$$\sigma^2 = \frac{1}{M}\sum_{j=1}^{M}\sigma_j^2 \tag{2.41}$$

Therefore, the coding gain is equal to

$$G_{TC} = \frac{\frac{1}{M}\sum_{j=1}^{M}\sigma_j^2}{\left(\prod_{j=1}^{M}\sigma_j^2\right)^{1/M}} \qquad (2.42)$$

The coding gain equals to the ratio of the arithmetic average over the geometric average of the variances of the transform coefficients. G_{TC} is maximized when KLT is used [23]. For a white Gaussian source, no coding gain is obtained since the variances of the transform coefficients are equal. Therefore, the transform coding is most beneficial for the correlated sources.

Adaptive Transform Coding (ATC) varies the bit allocation among the transform coefficients adaptively from frame to frame while keeping the total number of bits constant. This dynamic bit allocation is controlled by time-varying statistics of the speech signal which is transmitted as side information. The side information is also used to determine the step size of the various coefficient quantizers. Very good speech quality can be achieved at 12-16 kbits/s by ATC.

Chapter 3

The Vector Transform Quantization (VTQ) System

In this chapter, a vector transform quantization (VTQ) system is discussed. In this 2.4 kbps coding system, DCT is used as a transform operation and the split VQ is used for quantization of the transform coefficients. Furthermore, the transform coefficients are encoded by assigning more bits to more important transform coefficients. The bit assignment is optimized based on the asymptotic theory of the average energy distribution of the transform coefficients. In Section 3.1, the VTQ system is described. The optimal bit allocation strategy is discussed in Section 3.2.

3.1 System Description

Traditionally, scalar quantization is used to quantize the transform coefficients in the transform coding system. However, it is easy to show [27] that a better performance can be achieved by employing VQ rather than the scalar quantization even if the transform coefficients are not correlated. If there exists correlation among the transform coefficients, the VQ can further exploit this redundancy. Vector Transform Quantization (VTQ) coding system was developed as an effort to improve the performance of the transform coding system [27][45]. As we know, the codebook vectors in VQ have to be chosen such that they are representatives of the set of transform coefficients. However, a considerable amount of memory is required for storage of the VQ codebook. Also, the quantization of a vector requires a large number of calculations. Due to these considerations, a split VQ is usually used in VTQ system, where a set of m VQ's ($m<<M$), instead of one M-dimensional VQ, are employed. Obviously, in order to make the m VQ more efficient, an optimum bit assignment rule for split VQ is required for assigning more bits to more important transform coefficients than to less important coefficients.

The DCT is usually used in the VTQ system as a suboptimal orthogonal transform, although the optimal transform is the KLT which fully decorrelates the transform coefficients. A very low correlation between transform coefficients is achieved by the DCT as well. By choosing this transform, we avoid the problems of a signal dependent transform encoding, the determination of the correlations, and the computation of the

eigenvalues and eigenvectors. In addition, we can use a fast algorithm to compute the transform coefficients.

VTQ based on the DCT was originally proposed in [27][45] for speech coding at 8 kbps. Here, we re-design a VTQ system for coding at 2.4 kbps. Fig 3.1 illustrates the structure of the VTQ system. It is a coding system where each consecutive M samples of a waveform are transformed into a set of coefficients which is quantized by a set of $m<<M$ vector quantizers. An inverse vector quantization and inverse transform are taken at the decoder to obtain the corresponding block of reconstructed speech samples. The design of this system includes the choice of the transform T and the optimum bit assignment for coding the transform coefficients.

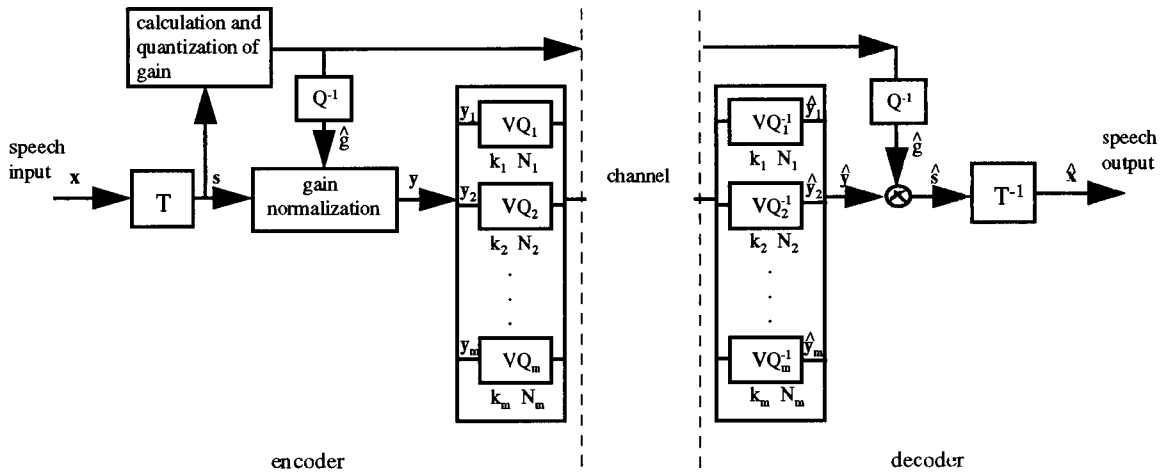


Fig 3.1 The block diagram of a typical VTQ system

Let $\mathbf{x}$ be a vector representing M consecutive samples of a speech subframe, $\mathbf{x}=[x_1,x_2,...,x_M]^T$, and $\mathbf{T}$ be the orthogonal transform DCT. Then, the transformed vector $\mathbf{s}$, $\mathbf{s}=[s_1,s_2,...,s_M]^T$, is given by

$$\mathbf{s} = \mathbf{T}\mathbf{x} \tag{3.1}$$

where $\mathbf{x}$ is assumed to have zero mean and $\mathbf{T}$ can be denoted by

$$\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_M]^T \tag{3.2}$$

$\mathbf{t}_i^T$, $i = 1,2,\ldots,M$, is the i th row vector of the transform matrix $\mathbf{T}$. As described in Chapter 2, $\mathbf{T}$ is an $M \times M$ matrix whose i th row, j th column element t_{ij} is as follows:

$$t_{ij} = \alpha(i)\cos\left[(2j-1)\frac{\pi(i-1)}{2M}\right] \qquad i, j = 1,2,\ldots,M \tag{3.3}$$

where $\alpha(1) = 1$ and $\alpha(i) = \sqrt{2}$ for $i = 2,\ldots,M$.

The transform vector $\mathbf{s}$ needs to be normalized before the quantization. Here, the gain g is chosen to be the largest value among transform coefficients of the two subframes which form a frame. Each frame has total 256 speech samples, correspondingly, each subframe has 128 samples. Scalar quantization is used to quantize the gain. The vector $\mathbf{s}$ is gain-normalized to a vector $\mathbf{y}$ with elements

$$y_i = \frac{s_i}{\hat{g}}, \qquad i = 1,2,\ldots,M \tag{3.4}$$

where $\hat{g}$ denotes the quantized gain. The normalized vector $\mathbf{y}$ is quantized to $\hat{\mathbf{y}}$ by m VQs with dimension k_i and codebook size N_i, $i = 1,2,\ldots,m$. The codebook vectors are stored in the gain-normalized form. The input of VQ_1 is composed of the first k_1 components of the vector $\mathbf{y}$ and is denoted by $\mathbf{y}_1$, the input of VQ_2 is the following k_2 components and is denoted by $\mathbf{y}_2$, and so on. At the decoder, the inverse quantizers

VQ_i^{-1}, $i = 1,2,...,m$, are applied and the corresponding quantized vectors $\hat{\mathbf{y}}_i$, $i = 1,2,...,m$, are retrieved from the codebooks. The concatenation, $\hat{\mathbf{y}}$, of vectors $\hat{\mathbf{y}}_i$, $i = 1,2,...,m$, is multiplied by $\hat{g}$ to obtain the quantized vector $\hat{\mathbf{s}}$. By taking the inverse transform $\mathbf{T}^{-1}$ to $\hat{\mathbf{s}}$, we can obtain the reconstructed speech vector $\hat{\mathbf{x}}$.

Our design target is the minimization of the distortion between the input vector $\mathbf{x}$ and the reconstructed vector $\hat{\mathbf{x}}$ at reasonable complexity. This distortion is measured by:

$$\begin{aligned} \varepsilon &= E[(\mathbf{x}-\hat{\mathbf{x}})^T(\mathbf{x}-\hat{\mathbf{x}})] \\ &= E[(\mathbf{s}-\hat{\mathbf{s}})^T(\mathbf{s}-\hat{\mathbf{s}})] \\ &= \hat{g}^2 E[(\mathbf{y}-\hat{\mathbf{y}})^T(\mathbf{y}-\hat{\mathbf{y}})] \end{aligned} \qquad (3.5)$$

Equation (3.5) indicates that to minimize the mean-squared error between $\mathbf{x}$ and $\hat{\mathbf{x}}$ is equivalent to minimize the mean-squared quantization error of $\mathbf{y}$.

Let ε_i be the MSE distortion of the quantizer VQ_i, then

$$\varepsilon_i = E\left[(\mathbf{y}_i-\hat{\mathbf{y}}_i)^T(\mathbf{y}_i-\hat{\mathbf{y}}_i)\right] \qquad (3.6)$$

then, equation (3.5) can be written as

$$\varepsilon = \hat{g}^2 E\left[(\mathbf{y}-\hat{\mathbf{y}})^T(\mathbf{y}-\hat{\mathbf{y}})\right] = \hat{g}^2 \sum_{i=1}^{m} \varepsilon_i \qquad (3.7)$$

Equation (3.7) shows that, for the given bit allocation, if the encoding process minimizes each distortion ε_i, the total distortion ε is minimized. In this system, the encoding in each of the quantizers VQ_i is done according to the minimum distortion criterion, and a full codebook search procedure is applied. In other words, for each vector $\mathbf{y}_i$, the

"nearest" vector $\hat{\mathbf{y}}_i$ is retrieved from the corresponding codebook C_i to minimize the distortion

$$\|\mathbf{y}_i - \hat{\mathbf{y}}_i\|^2 = \min_{\mathbf{u}\in C_i} \|\mathbf{y}_i - \mathbf{u}\|^2 \qquad i = 1,2,\ldots,m \qquad (3.8)$$

In order to minimize the MSE between the source and the reproduced sequence, we will discuss a bit assignment rule for finding the optimal bit allocation for the VQ of the transform coefficients in Section 3.2.

3.2 Bit Allocation Optimization for Vector Quantization

The input speech signal sampled at 8 kHz is segmented in frames of 256 samples. The data in a frame are encoded with a fixed number of bits. At the rate of 2.4 kbps, the total number of bits for a frame is 77. Each frame is divided into two subframes of dimension $M = 128$. We assign 5 bits for the quantization of the gain for each frame and 36 bits for the VQ of the transform coefficients for each subframe.

Assuming that the M-dimensional vector $\mathbf{y}$ is quantized by a set of m VQ's having dimension k_i and codebook size N_i, $i = 1,2,\ldots,m$, then

$$M = \sum_{i=1}^{m} k_i \qquad (3.9)$$

The average rate r in bits/sample is

$$r = \frac{1}{M}\sum_{i=1}^{m} \log_2 N_i = \frac{B_T}{M} \qquad (3.10)$$

where B_T is the number of code bits available for the vector $\mathbf{y}$. Here, $B_T = 36$. Following a derivation similar to [27], the strategy of the optimal bit assignment for VQ's can be summarized to the following steps:

1. Compute the estimated variances, $\hat{\sigma}_i^2$, of transform coefficients.

$$\hat{\sigma}_i^2 = \frac{1}{n}\sum_{j=1}^{n} y_{ij}^2 \qquad i = 1,2...M \qquad (3.11)$$

where $\hat{\sigma}_i^2$ is the estimated variance of the ith component in the vector y, n is the number of vectors which are used for estimating statistics, and y_{ij} is the value of the ith component in the jth vector, $i = 1,2...M$, $j = 1,2...n$. The input vectors are assumed to have zero mean.

2. Compute the number of bits, b_i assigned to the ith sample [23][27]

$$b_i = r + \beta_i + \frac{1}{2}\log_2 \frac{\hat{\sigma}_i^2}{(\prod_{j=1}^{M} \hat{\sigma}_j^2)^{\frac{1}{M}}} \qquad i = 1,2...M \qquad (3.12)$$

where β_i is a term which depends on quantization coefficients and on vector dimensions. Here, we assume $\beta_i = 0$.

3. If $b_i < 0$ for any of $i = 1,2...M$, set $b_i = 0$. Then, the remaining components with a positive bit allocation are subject to bit reassignment. The reoptimization algorithms have been proposed in [23][46][47].

4. Calculate the number of code bits, B_1, which is allocated to the VQ_1

$$B_1 = \sum_{i=1}^{k_1} b_i \tag{3.13}$$

The choice of the dimension k_i is guided by implementation constraints. Beginning with k_1, each dimension is set to the maximal value which can be implemented under a given complexity constraint, except k_m which is chosen to satisfy (3.9). The complexity constraint is actually a limitation on the codebook size

$$0 < B_i < \log_2 N_{\max}, \qquad i = 1,2,...,m \tag{3.14}$$

where $N_{\max}$ is the maximal codebook size and set to be 1024.

5. Similar to step 4, choose k_2 and obtain B_2; choose k_3 and obtain B_3 and so on.

$$B_2 = \sum_{i=k_1+1}^{k_1+k_2} b_i, \quad B_3 = \sum_{i=k_1+k_2+1}^{k_1+k_2+k_3} b_i, \quad ....., B_m = \sum_{i=k_1+...+k_{m-1}+1}^{k_1+...+k_m} b_i \tag{3.15}$$

6. Round off each B_i $(i = 1,2...m)$ to its nearest integer value and adjust them to satisfy

$$B_T = \sum_{i=1}^{m} B_i \tag{3.16}$$

Chapter 4

Transform Coding with the Coefficient Ranking

In this chapter, two transform coding systems with coefficient ranking are introduced for speech coding at 2.4 kbps. In Section 4.1, a new Vector Transform Quantization with the Coefficient Ranking (VTQ-CR) system is proposed, in which the transform coefficients are ranked and a part of them with higher energies are vector quantized and transmitted. Based on the VTQ-CR system, another transform coding system, Vector Transform Quantization with the Coefficient Ranking and Adaptive Linear Prediction (VTQ-CR-ALP), is introduced in Section 4.2. An adaptive transform domain linear predictor is applied to reduce the near-sample correlations of the ranked transform coefficients and the analysis-by-synthesis technique is used to determine the optimal excitation of the synthesis filter.

4.1 The VTQ with the Coefficient Ranking (VTQ-CR) System

4.1.1 Ranking Structure and Vector Quantization

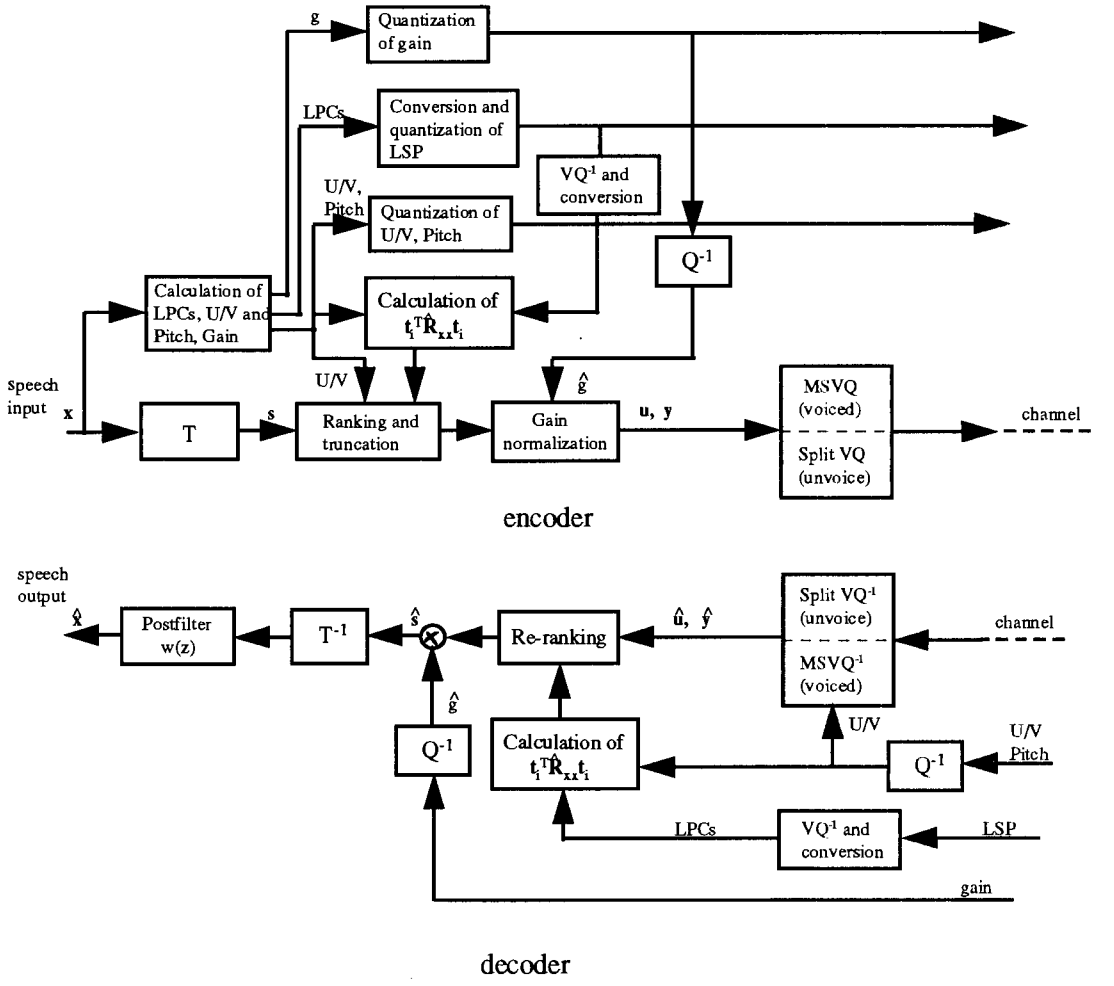
At low bit rates, it is important to maximize the cost effectiveness of every bit that is transmitted. The objective here is to get the best approximation of input speech in the sense of minimization of MSE between the original speech and the reconstructed waveform. The coefficient ranking scheme introduced here can improve the efficiency of the VQ.

In the transform coding systems, the signal power is distributed non-uniformly over the transform coefficients. Some of the coefficients can be set to zero without introducing any perceptual distortion because the corresponding basis components can not be observed and some can be quantized less accurately because the human observer is not very sensitive to errors. For other components, people may have great sensitivity to errors, which means that the corresponding coefficients should be quantized more accurately. The idea of the transform coefficient ranking is that, based on the short-time spectral information of the input speech, the transform coefficients are ranked in a descending order of their energy values. Only the first N_t ranked coefficients with higher energy values are vector quantized, while others with lower energy values are set to zero.

Ranking the transform coefficients and discarding those with lower energy values can make the VQ more efficient.

4.1.2 System Description

A VTQ System with Coefficient Ranking (VTQ-CR) was developed for the goal of improving the efficiency of the VQ and the system performance. Fig 4.1 is a block diagram of the VTQ-CR system.



LPCs: Linear Prediction Coefficients

Fig 4.1 The block diagram of a VTQ-CR system

The speech signal is segmented in frames of 256 samples and encoded with a fixed number of bits per frame. Each frame is divided into two subframes of 128 samples each, i.e., $M = 128$, where M denotes the length of a subframe. The transform vector, **s**, $\mathbf{s} = [s_1, s_2, \ldots, s_M]^T$, is obtained by

$$\mathbf{s} = \mathbf{T}\mathbf{x} \tag{4.1}$$

where **x** is an input subframe vector consisting of M consecutive samples, $\mathbf{x} = [x_1, x_2, \ldots, x_M]^T$, and **T** is the DCT orthogonal transform matrix. Its elements are defined as in (2.28). The transform coefficients, $s_1, s_2, \ldots, s_M$, are ranked in a descending order of their energy values to form a new vector. It means that the variances of the transform coefficients are in order such that the successive values contribute proportionally less and less to the total. The variances represent the energy or information content of the corresponding transform coefficients. Only the N_t most significant coefficients are vector quantized and transmitted, while others are set to zero. After the same gain-normalization operation as in the VTQ system, the truncated representation of the new ranked transform vector is given by

$$\mathbf{u} = [y_1, y_2, \ldots, y_{N_t}, 0, \ldots, 0]^T \tag{4.2}$$

where $y_i = s_j / \hat{g}$, $i = 1,2,\ldots,N_t$, $j \in [1, M]$, and $\hat{g}$ is the quantized gain, which means that the jth normalized transform coefficient is ranked as the ith component of the new formed vector **u**. The truncated vector is denoted by **y**

$$\mathbf{y} = \left[y_1, y_2, \ldots, y_{N_t} \right]^T \tag{4.3}$$

The truncation error is given by the sum of the variances of the discarded coefficients.

Fig 4.2 and Fig 4.5 show segments of voiced speech signal and unvoiced speech signal, respectively. Fig 4.3, Fig 4.4 and Fig 4.6, Fig 4.7 show the corresponding non-ranked and ranked transform coefficients for the voiced and the unvoiced speech signal, respectively.

From Fig 4.3 and Fig 4.6, we can see that the energy distribution over the non-ranked transform coefficients is non-uniform and the coefficients with greater variances may occur anywhere. With very limited code bits at the bit rate of 2.4 kbps, the VTQ coding scheme discussed in Chapter 3 could not quantize these significant coefficients effectively because Equation (3.11) is used to compute the variances of the transform coefficients in the optimal bit allocation under the assumption that the input signal is a stationary process. Speech signal is non-stationary over a long interval of time, although it can be considered stationary over a sufficiently short time interval. However, by applying the coefficient ranking technique in the VTQ-CR system, these coefficients are concentrated in the first part of the vector and ranked in a descending order of their energy values as shown in Fig 4.4 and Fig 4.7.

The ranking information of the transform coefficients is extracted from the speech spectral information. The variances of the transform coefficients are calculated from the basis vectors of the transform matrix **T** and the short-time autocorrelation matrix of the input speech, which can be obtained from the linear prediction coefficients. Therefore, in addition to the transmission of the quantized truncated transform vector, the linear

prediction coefficients have to be computed once per frame and transmitted in the form of quantized LSPs.

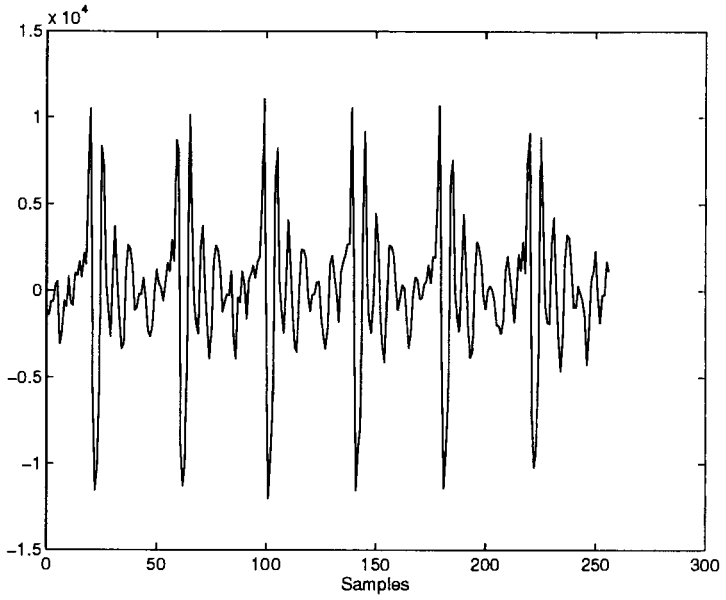


Fig 4.2 Voiced speech signal

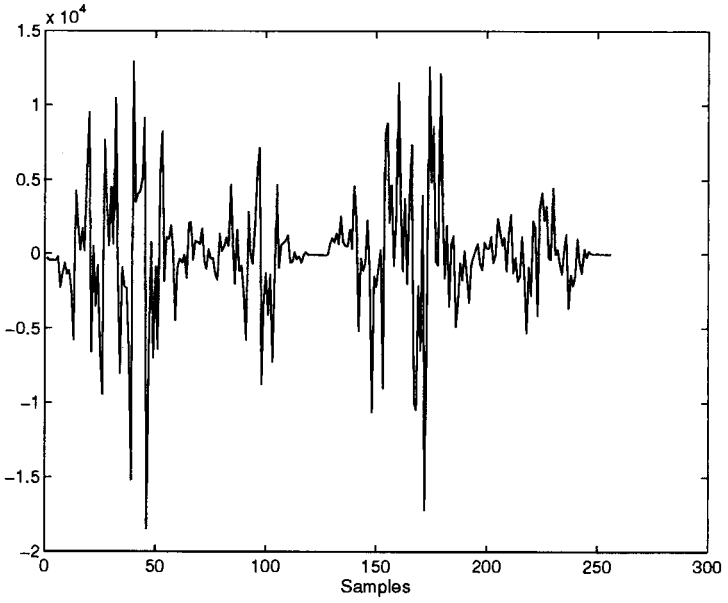


Fig 4.3 Transform coefficients of the voiced speech signal

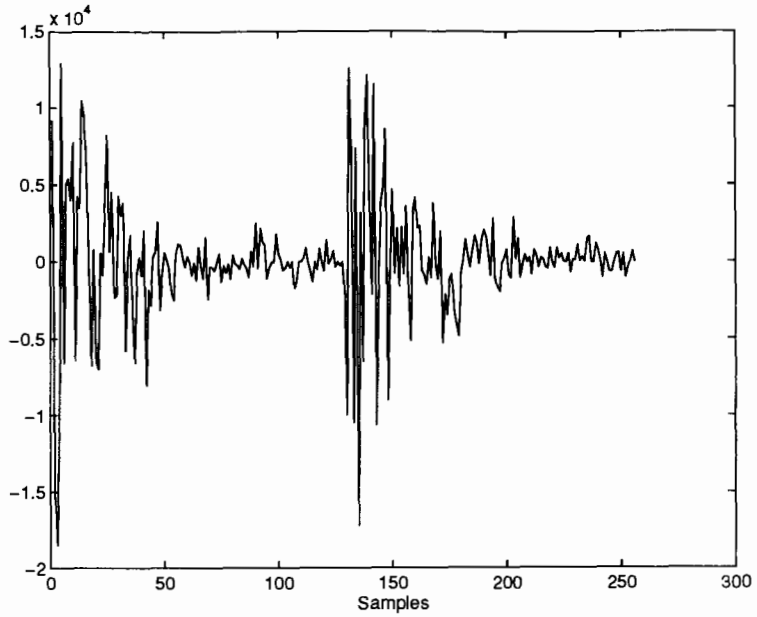


Fig 4.4 Ranked transform coefficients of the voiced speech signal

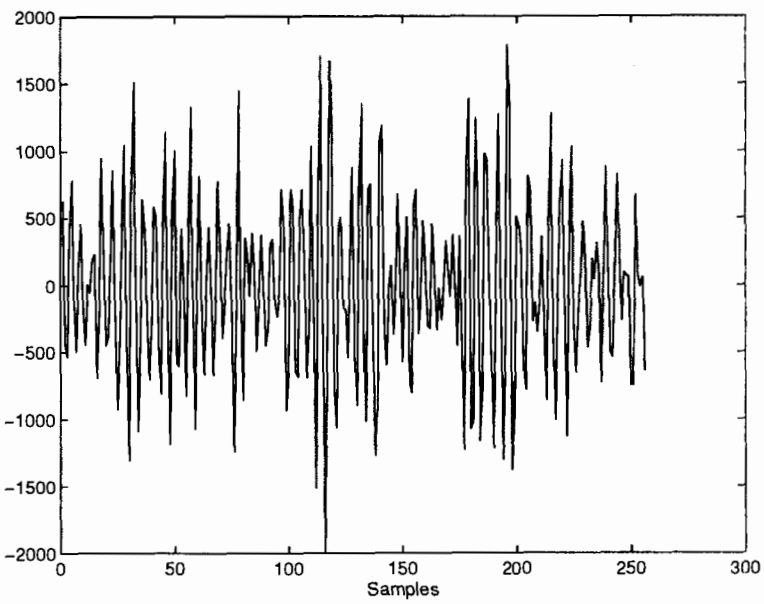


Fig 4.5 Unvoiced speech signal

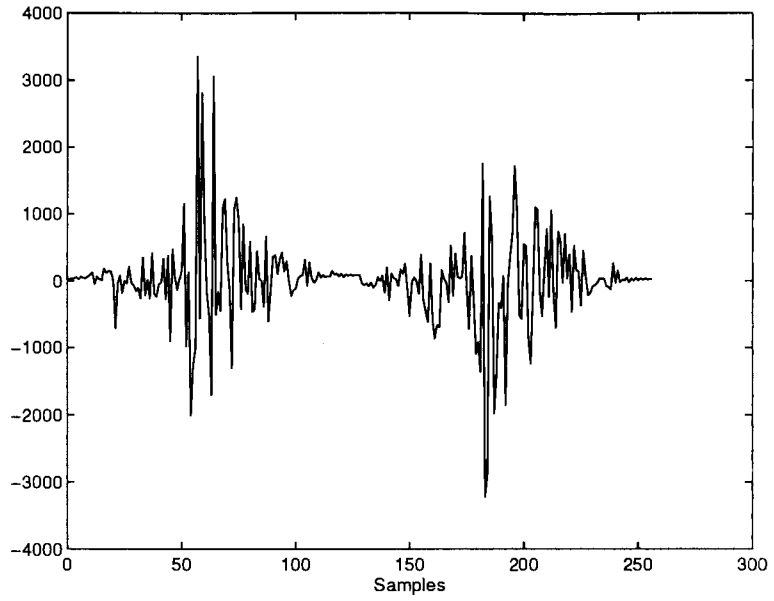


Fig 4.6 Transform coefficients of the unvoiced speech signal

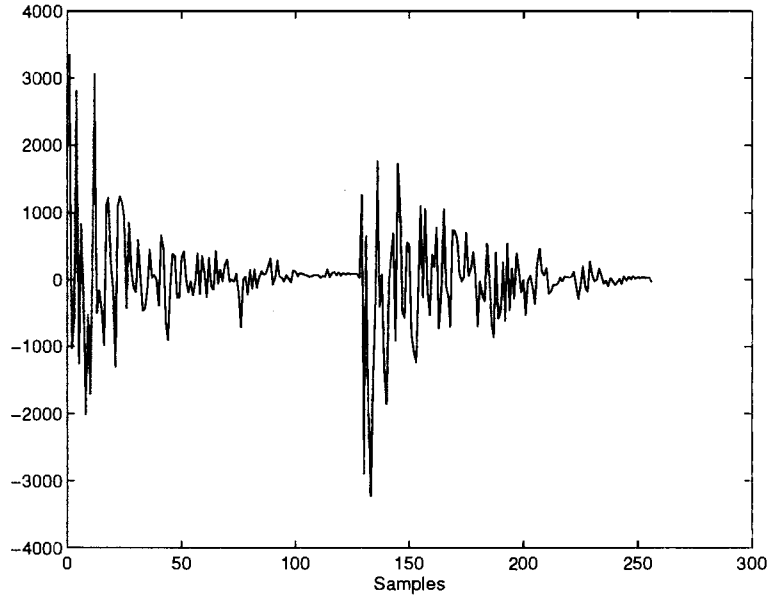


Fig 4.7 Ranked transform coefficients of the unvoiced speech signal

As discussed in Chapter 2, the linear prediction coefficients are converted to LSPs for quantization. The LSPs are determined by using the method proposed by Kabal [48]. Split VQ is applied for the quantization of the LSPs.

Voiced speech and unvoiced speech have different characteristics. As shown in Fig 4.2 and Fig 4.5, voiced speech waveform is periodic at a rate corresponding to the glottal pulse frequency, while unvoiced signal is random noise-like waveform. Compared with voiced frames, the ranked transform coefficients of unvoiced frames have much lower energy level and more scattered energy distribution as shown in Fig 4.4 and Fig 4.7. This indicates that more coefficients with low energy values could be discarded without leading to a great perceptual distortion in the case of voiced speech than in the case of unvoiced speech. In order to make the encoding more efficient, voiced and unvoiced frames are classified. Different truncations are taken for the voiced and unvoiced transform vectors; and different VQ codebooks are searched for the quantizations of the gains and the ranked transform coefficients. The pitch information is used to determine the period of the autocorrelation function for voiced speech. The pitch periods of the voiced frames, the voicing decision and the gain are transmitted once per frame.

At the encoder, a two-stage VQ is applied for the quantization of the voiced truncated transform vectors. The first stage performs a relatively crude quantization and then the second one provides a further refinement. For the unvoiced truncated transform vectors, split VQ is used.

At the decoder, an inverse VQ is taken and the quantized truncated vector is obtained as $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{N_t}]^T$. The corresponding ranked transform coefficients are reconstructed as $\hat{\mathbf{u}}$, $\hat{\mathbf{u}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{N_t}, 0, \ldots, 0]^T$. Based on the linear prediction coefficients, the pitch period and the voicing decision, the components of the vector $\hat{\mathbf{u}}$ are re-ranked to restore the order of the coefficient sequence. The re-ranked sequence is multiplied by the quantized gain, $\hat{g}$, to obtain the quantized transform vector $\hat{\mathbf{s}}$. Then, an inverse DCT transform is taken to reconstruct the corresponding block of samples $\hat{\mathbf{x}}$. To further enhance the perceptual quality of the reconstructed speech, a postfilter is added to the decoder output [49].

4.1.3 Coefficient Ranking

At the encoder, the transform coefficients, $s_1, s_2, \ldots, s_M$, are ranked in a descending order of their energy values or variances. Assuming that the input speech is zero mean, the variance of the transform coefficient s_i, σ_i^2, is defined as

$$\sigma_i^2 = E[s_i^2] \tag{4.4}$$

From the equation (4.1), the *i*th component of the transform vector $\mathbf{s}$ is given by

$$s_i = \mathbf{t}_i^T \mathbf{x} \tag{4.5}$$

where $\mathbf{s} = [s_1, s_2, \ldots, s_M]^T$, and $\mathbf{t}_i^T$ is the i th row vector of the transform matrix $\mathbf{T}$, $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_M]^T$. Hence, the equation (4.4) becomes

$$\begin{aligned}\sigma_i^2 = E[s_i^2] &= \mathbf{t}_i^T E[\mathbf{x}\mathbf{x}^T]\mathbf{t}_i \\ &= \mathbf{t}_i^T \mathbf{R}_{xx} \mathbf{t}_i \\ &\approx \mathbf{t}_i^T \hat{\mathbf{R}}_{xx} \mathbf{t}_i\end{aligned} \tag{4.6}$$

where $\hat{\mathbf{R}}_{xx}$ is an approximation of the $M \times M$ autocorrelation matrix, $\mathbf{R}_{xx}$, of the input speech frame, which is obtained from the linear prediction coefficients and is updated at intervals of a frame of the speech signal. If the approximations of the corresponding autocorrelation functions $r_{xx}(n)$, $n = 0,1,2,\ldots,M-1$ are denoted by $\hat{r}_{xx}(0), \hat{r}_{xx}(1), \ldots, \hat{r}_{xx}(M-1)$,

$$\hat{\mathbf{R}}_{xx} = \begin{bmatrix} \hat{r}_{xx}(0) & \hat{r}_{xx}(1) & \hat{r}_{xx}(2) & \cdots & \hat{r}_{xx}(M-1) \\ \hat{r}_{xx}(1) & \hat{r}_{xx}(0) & \hat{r}_{xx}(1) & \cdots & \hat{r}_{xx}(M-2) \\ \hat{r}_{xx}(2) & \hat{r}_{xx}(1) & \hat{r}_{xx}(0) & \cdots & \hat{r}_{xx}(M-3) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \hat{r}_{xx}(M-1) & \hat{r}_{xx}(M-2) & \hat{r}_{xx}(M-3) & \cdots & \hat{r}_{xx}(0) \end{bmatrix} \tag{4.7}$$

We will discuss the calculation procedure for the $\hat{\mathbf{R}}_{xx}$ in the Subsection 4.1.4.

Direct computation of σ_i^2 using Equation (4.6) will involve a significant amount of multiplication operations. There are $M^2 + M$ multiplications for the computation of each σ_i^2. However, since the vector t_i is a column vector of a DCT matrix, we can resort to the fast DCT algorithm to reduce the computation complexity of σ_i^2. Let the first two terms in (4.6) be denoted by $\mathbf{c}_i = \mathbf{t}_i^T \hat{\mathbf{R}}_{xx}$, where $\mathbf{c}_i$ is a row vector. In the fast DCT, the computation complexity for the transform $\mathbf{s} = \mathbf{T}\mathbf{x}$ is $M/2 \log_2 M$, where $\mathbf{T}$ is a $M \times M$ DCT square matrix and $\mathbf{x}$ is the input vector. Suppose that we take M DCT for different

input vectors, the complexity would be $M\left(M/2 \log_2 M\right)$ which is equal to the computation complexity of M $\mathbf{c}_i$'s for different i. Also, there are M multiplications for each $\mathbf{c}_i\mathbf{t}_i$. For a frame of speech samples, we need compute 128 variances of the transform coefficients. Therefore, the total number of multiplication operations is $M\left(M/2 \log_2 M + M\right) = 73728$ compared to $M^3 + M^2 = 2113536$ for a direct use of (4.6).

4.1.4 Autocorrelation Functions and Linear Prediction Coefficients

We assume that a frame of $2M$ consecutive samples of the speech signal is a realization of a stationary process. The autocorrelation functions are estimated, based on a frame of $2M$ samples, in the encoder by using the autocorrelation method described in Chapter 2. We rewrite the equation (2.15) as follows:

$$r_{xx}(n) = \frac{1}{2M} \sum_{i=1}^{2M-n} x_i w_i x_{n+i} w_{n+i} \qquad n = 0,1,2,\ldots,2M-1 \qquad (4.8)$$

where w_i is the Hamming window function. By knowing the $r_{xx}(n)$'s, $n = 0,1,2,\ldots,16$, a 16th order LPC analysis is carried out in the encoder for each frame. The linear prediction coefficients a_i, $i = 1,2,\ldots,16$, can be determined from the Yule-Walker equation in (2.10).

In order to use the identical autocorrelation matrix $\hat{\mathbf{R}}_{xx}$, both the encoder and the decoder calculate the autocorrelation functions, $\hat{r}_{xx}(n)$'s, $n = 0,1,\ldots,M-1$, based on the knowledge of the linear prediction coefficients which are vector quantized in the line

spectral frequency domain [38][39]. $\hat{r}_{xx}(0)$ is normalized to be 1. The inverse Levinson-Durbin procedure is used to convert the quantized linear prediction coefficients $\hat{a}_i$, $i = 1,2,\ldots,16$, to the autocorrelation functions $\hat{r}_{xx}(n)$, $n = 1,2,\ldots,16$ [12].

Different algorithms are applied to estimate $\hat{r}_{xx}(n)$, $n = 17,18,\ldots, M-1$, depending on whether the speech frame is voiced or unvoiced. For a voiced speech frame, the short-time autocorrelation function exhibits peaks at time-shifts corresponding to multiples of the pitch period P which is considered as the period of the short-time autocorrelation function. We define int(a) as an integer operator of a. When $\text{int}(P/2) > 16$, $\hat{r}_{xx}(n)$, $n = 17,\ldots,\text{int}(P/2)$, can be extrapolated by the following equation:

$$\hat{r}_{xx}(n) = \sum_{i=1}^{16} \hat{a}_i \hat{r}_{xx}(n-i) \qquad 16 < n \le \text{int}(P/2) \qquad (4.9)$$

$\hat{r}_{xx}(n)$, $\max(16, \text{int}(P/2)) < n \le P$, can be obtained by taking advantage of the symmetry of the autocorrelation function within one period.

$$\hat{r}_{xx}(n) = \hat{r}_{xx}(P-n) \qquad \max(16, \text{int}(P/2)) < n \le P \qquad (4.10)$$

Because of the periodicity, $\hat{r}_{xx}(n)$, $P < n \le M-1$, can be obtained by duplicating the values in the first period. Fig 4.8 shows the extrapolated autocorrelation function of a voiced speech frame.

Unvoiced speech frames, on the other hand, do not usually have strong short-term correlations. Therefore, the corresponding autocorrelation functions beyond (0,16) can be extrapolated by

$$\hat{r}_{xx}(n) = \sum_{i=1}^{16} \hat{a}_i \hat{r}_{xx}(n-i) \qquad 16 < n \leq M-1 \qquad (4.11)$$

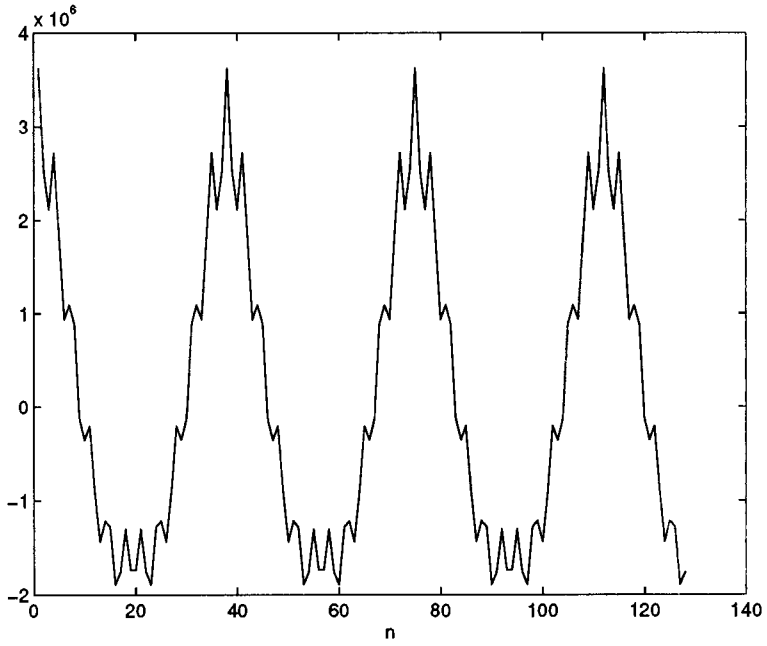


Fig 4.8 Extrapolated autocorrelation function of a voiced frame (P=37)

4.1.5 Voiced/Unvoiced Classification and Pitch Extraction Algorithm

The pitch period can be determined from the autocorrelation function, $r_{xx}(n)$, $n = 0,1,\ldots,2M-1$, of the input speech frame. In the case of voiced speech, the main peak in the short-time autocorrelation function normally occurs at a lag equal to the pitch

period. This peak is therefore detected and its time position gives the pitch period, P, of the input speech. In the case of unvoiced speech, the short-time autocorrelation function exhibits no strong peaks and overall has a much lower amplitude. Fig 4.9 and Fig 4.10 show plots of the short-time autocorrelation functions for voiced and unvoiced speech, respectively.

Based primarily on the strength of the pitch periodicity of the short-time autocorrelation function, the voicing decision is made on the ratio of the amplitude of the main peak to the amplitude of the autocorrelation function for zero time lag, that is $r_{xx}(0)$. The threshold is set to be 0.35. If the amplitude of the main peak is less than $0.35r_{xx}(0)$, the speech frame is declared unvoiced, otherwise it is voiced.

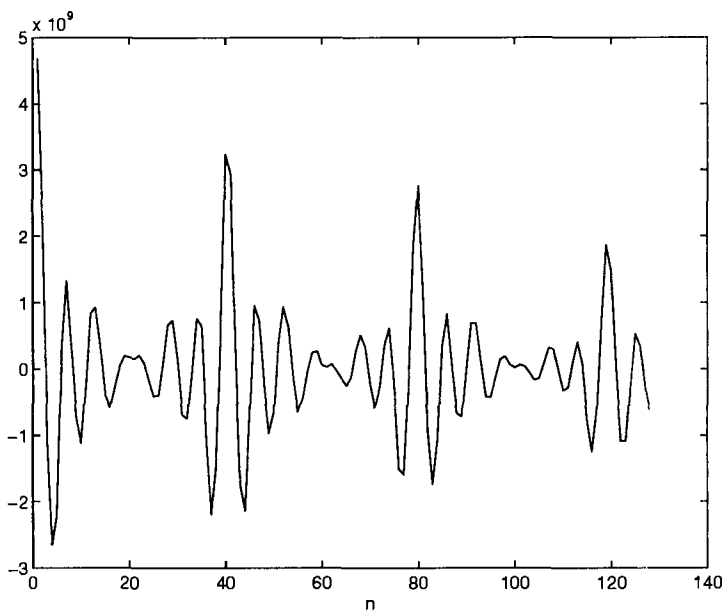


Fig 4.9 Short-time autocorrelation function for voiced speech

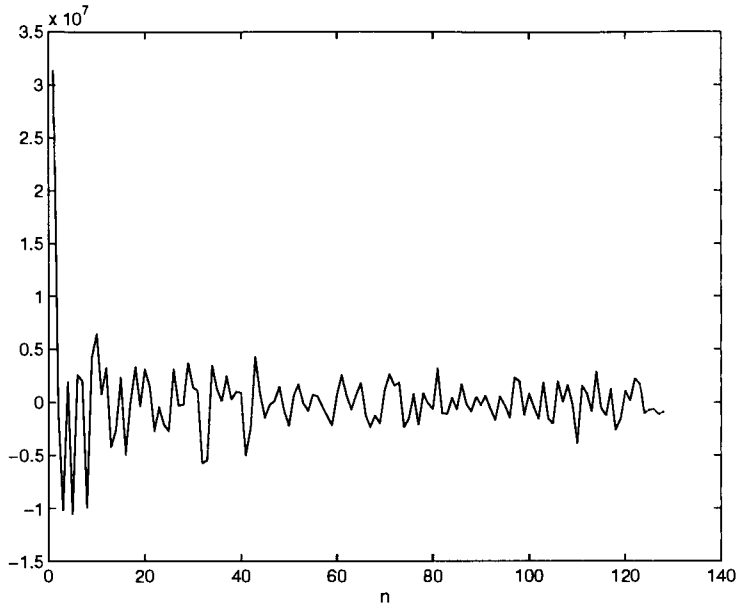


Fig 4.10 Short-time autocorrelation function for unvoiced speech

4.1.6 Vector Quantization of the Ranked Transform Coefficients

A two-stage vector quantizer is used for the VQ of the voiced truncated transform vectors. The structure for the two-stage VQ is depicted in Fig 4.11. For each voiced truncated transform vector, the available B_T code bits are divided with B_1 bits for the first stage and B_2 bits for the second stage. The truncated vector $\mathbf{y}$ is quantized with the first stage codebook producing the selected first-stage code vector $\hat{\mathbf{y}}_1$. An error vector $\mathbf{E}_1$ is formed by subtracting $\hat{\mathbf{y}}_1$ from $\mathbf{y}$. Then $\mathbf{E}_1$ is quantized using the second stage codebook with exactly the same procedure as in the first stage and the selected second-stage code vector $\hat{\mathbf{y}}_2$ is produced. The decoder receives for each stage an index identifying the stage

code vector selected and forms the reproduction $\hat{\mathbf{y}}$ by summing the vectors $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$, i.e.,

$$\hat{\mathbf{y}} = \hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2 \tag{4.12}$$

The overall quantization error $\mathbf{y} - \hat{\mathbf{y}}$ is equal to the quantization residual from the second stage. Sequential searching of the stage codebooks renders the encoding complexity $2^{B_1} + 2^{B_2}$.

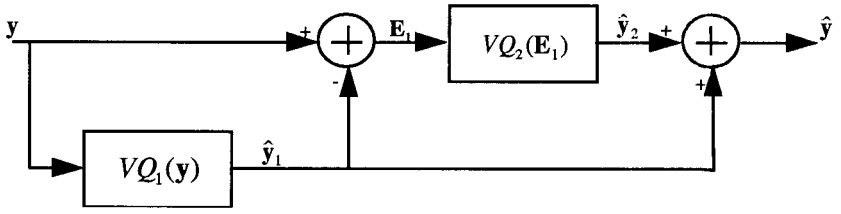


Fig 4.11 Two-stages VQ in VTQ-CR system

The conventional way of designing the stage codebooks for sequential search MSVQ with a MSE distortion criterion is to apply the generalized Lloyd algorithm (GLA) stage-by-stage for a given training set of input vectors. Here, the training data for the first stage VQ codebook consists of the truncated transform vectors. The codebooks are trained sequentially, the second stage using a training sequence consisting of quantization error vectors from the first stage.

For the unvoiced truncated transform vectors, split VQ is used as in the VTQ system, but no optimum bit assignment algorithm is involved. The bit allocation is guided by comprehensive consideration of such factors as the energy concentration of the transform coefficients, the complexity of codebook search and the efficiency of the VQ.

4.1.7 Bit Allocation

At the rate of 2.4 kpbs, as analyzed in Chapter 3, a frame of 256 samples is coded by 77 bits. Each frame consists of two subframes with each dimension 128. Two classes of VQ codebooks are designed to encode the voiced and unvoiced speech, respectively. The bit allocations for voiced and unvoiced frames are as follows:

	Voiced (bits)	Unvoiced (bits)
LSP	24	24
U/V	1	1
Pitch	7	0
Gain	5	7
Transform Vector	2*20	2*22

Table 4.1 The bit allocation for VTQ system

The detailed truncation strategies and bit allocations for both voiced and unvoiced truncated transform vectors will be given in Chapter 5.

Linear prediction coefficients, voicing decision, pitch period and gain are updated per frame. The linear prediction coefficients are converted to the LSPs which are quantized by a split VQ. 16 LSPs are devided into three groups with first two having 5 for each and the last one having 6. With 24 bits available, a split VQ is designed with 8 bits assigned for each group.

For a voiced frame, 7 bits are required to quantize all possible 128 pitch lags. For the scalar quantization of the gain, 5 bits are used for a voiced frame and 7 bits are used for an unvoiced frame.

4.1.8 Postfiltering

The characteristics of human auditory perception is that speech formants are much more important to perception than spectral valleys. A postfilter [49] is added to the decoder output to enhance the perceptual quality of the reconstructed speech. The postfilter consists of a pole-zero filter based on the quantized short-term predictor coefficients followed by a spectral tilt compensator. The transfer function of the pole-zero filter is

$$W(z) = \frac{1-(\hat{a}_1\alpha\, z^{-1} + \hat{a}_2\alpha^2 z^{-2} + \cdots + \hat{a}_{16}\alpha^{16} z^{-16})}{1-(\hat{a}_1\beta\, z^{-1} + \hat{a}_2\beta^2 z^{-2} + \cdots + \hat{a}_{16}\beta^{16} z^{-16})} \qquad (4.13)$$

where $\hat{a}_i$, $i = 1,2,\ldots,16$, are the quantized linear prediction coefficients. $\alpha = 0.8$ and $\beta = 0.95$. An automatic gain control is used to avoid large gain excursions.

4.2 The Application of Linear Prediction in VTQ-CR---VTQ-CR-ALP

4.2.1 The Features of the VTQ-CR-ALP System

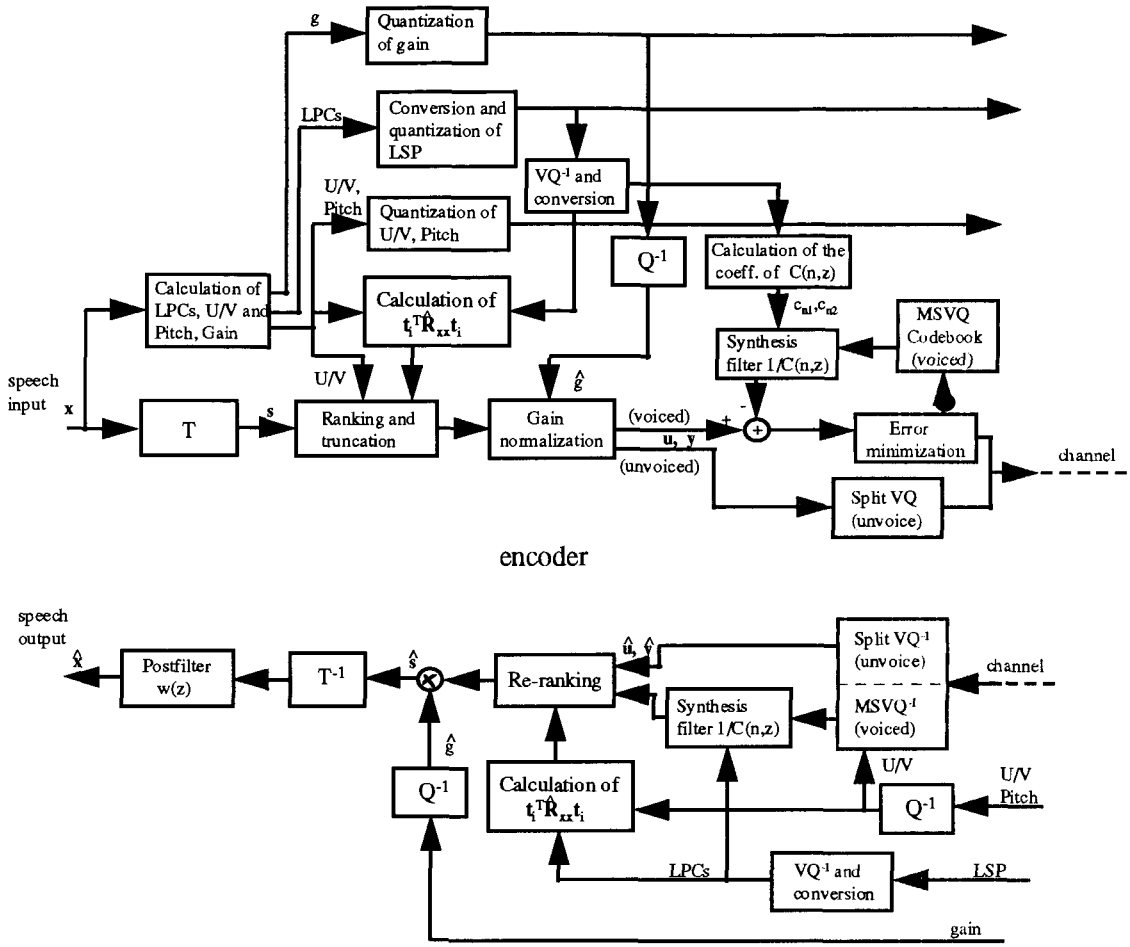
Although the DCT decorrelates the transform coefficients, there still exist correlations between the ranked transform coefficients, especially between the voiced ranked transform

coefficients. Hence, we propose another coder, Vector Transform Quantization with the Coefficient Ranking and Adaptive Linear Prediction (VTQ-CR-ALP), which is developed as an enhanced version of the VTQ-CR to improve the system performance further by exploiting possible redundancy between the voiced ranked transform coefficients. Fig 4.12 shows the structure of this system.

In common with VTQ-CR, VTQ-CR-ALP also makes use of the transform coefficient ranking technique. Moreover, after the gain normalization, a 2nd-order adaptive transform domain linear predictor is applied to the voiced ranked transform coefficients, which reduces their near-sample correlations. The analysis-by-synthesis method is introduced to determine an optimal excitation signal for reproducing the voiced truncated transform vectors. The transfer function of the linear predictor is given by

$$C(n,z) = 1 - c_{n1}z^{-1} - c_{n2}z^{-2} \qquad n = 1,2,\ldots,N_t \qquad (4.14)$$

where c_{n1} and c_{n2} are linear predictor parameters. They are chosen to minimize the MSE of the prediction residual which is obtained by filtering the voiced truncated transform vector $\mathbf{y}$ through the all-zero filter $C(n,z)$. The calculation of c_{n1} and c_{n2} only depends on the basis vectors of the DCT matrix $\mathbf{T}$ and the short-term speech spectral information which has to be transmitted for the coefficient ranking and re-ranking even in the case of no adaptive transform domain linear prediction. Therefore, there is no extra side information transmitted due to the application of the adaptive transform domain linear prediction. The derivation of the predictor parameters will be discussed in Subsection 4.2.2.



LPCs: Linear Prediction Coefficients

Fig 4.12 The block diagram of the VTQ-CR-ALP

The synthesized truncated transform vectors are produced by feeding an all-pole filter $1/C(n,z)$ with an excitation signal selected from a codebook. For a voiced speech subframe vector $\mathbf{x}$, letting the excitation of the corresponding truncated transform vector $\mathbf{y}$ be the vector $\mathbf{r}$ and the output of the filter $1/C(n,z)$ be $\hat{\mathbf{y}}$, then, the excitation vector, $\mathbf{r}$, should be selected such that the distortion ε_y between $\mathbf{y}$ and $\hat{\mathbf{y}}$ is minimized.

$$\varepsilon_{\mathbf{y}} = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \qquad (4.15)$$

The closed-loop derived excitation signal, $\mathbf{r}$, is quantized and the index of the corresponding codevector is transmitted to the decoder. The identical synthesis filter $1/C(n,z)$ is also used in the decoder to obtain the optimal synthesized signal $\hat{\mathbf{y}}$. Assuming that $\mathbf{r} = [r_1, r_2, \ldots, r_{N_t}]^T$, and $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{N_t}]^T$, the synthesized signal is given by

$$\hat{y}_n = c_{n1}\hat{y}_{n-1} + c_{n2}\hat{y}_{n-2} + r_n \qquad n = 1,2,\ldots,N_t \qquad (4.16)$$

where $\hat{y}_0 = \hat{y}_{-1} = 0$.

In this system, for the voiced subframes, a two-stage VQ is applied for the quantization of the excitation, **r**, of the synthesizer. For the unvoiced subframes, they are coded in the same way as in the VTQ-CR system.

4.2.2 Adaptive Transform Domain Linear Prediction Analysis

When the voiced ranked transform coefficients are filtered through the filter, $C(n,z)$, their near-sample correlations are reduced and a residual signal is produced as the output. The parameters, c_{n1}, c_{n2}, of the filter are determined by minimizing the energy of the residual signal. The derivation of these parameters is as follows:

For the ith input of the filter $C(n,z)$, y_i, $i \in [1, N_t]$, the transfer function is as follows

$$C(i,z) = 1 - c_{i1}z^{-1} - c_{i2}z^{-2} \tag{4.17}$$

The variance of the output residual signal is given by

$$\sigma_{e_i}^2 = E\left[\left\| y_i - c_{i1}y_{i-1} - c_{i2}y_{i-2} \right\|^2\right] \tag{4.18}$$

Let $\mathbf{C} = \left[\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_{N_t}\right]$, where $\mathbf{C}_i = \left[c_{i1}, c_{i2}\right]^T$ and $\mathbf{Y} = \left[\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_{N_t}\right]$, where $\mathbf{Y}_i = \left[y_{i-1}, y_{i-2}\right]^T$, then (4.18) can be written as

$$\sigma_{e_i}^2 = E\left[\left\| y_i - \mathbf{Y}_i^T\mathbf{C}_i \right\|^2\right] \tag{4.19}$$

In order to minimize $\sigma_{e_i}^2$, let $\partial\sigma_{e_i}^2 / \partial\,\mathbf{C}_i = 0$

Then, $$E\left[\mathbf{Y}_i\left(y_i - \mathbf{Y}_i^T\mathbf{C}_i\right)\right] = 0 \tag{4.20}$$

Let $\mathbf{Q} = \left[\mathbf{Q}_1, \mathbf{Q}_2, ..., \mathbf{Q}_{N_t}\right]$, where $\mathbf{Q}_i = E\left[\mathbf{Y}_i\mathbf{Y}_i^T\right]$ and $\mathbf{q} = \left[\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_{N_t}\right]$, where $\mathbf{q}_i = E\left[\mathbf{Y}_i y_i\right]$, then, the solution of the equation (4.20) is given by

$$\mathbf{C}_i = \mathbf{Q}_i^{-1}\mathbf{q}_i \tag{4.21}$$

Assuming that $y_i = s_j / \hat{g}$, $y_{i-1} = s_k / \hat{g}$, $y_{i-2} = s_h / \hat{g}$, $i = 1,2,...,N_t$, $j,k,h \in \left[1,2,...,M\right]$ and $y_{-1} = y_0 = 0$, based on the fact that $s_i = \mathbf{t}_i^T\mathbf{x}$ and $E\left[\mathbf{x}\mathbf{x}^T\right] \approx \hat{\mathbf{R}}_{xx}$, we obtain

$$\mathbf{Q}_i \approx \frac{1}{\hat{g}^2}\begin{bmatrix} \mathbf{t}_k^T\hat{\mathbf{R}}_{xx}\mathbf{t}_k & \mathbf{t}_h^T\hat{\mathbf{R}}_{xx}\mathbf{t}_k \\ \mathbf{t}_h^T\hat{\mathbf{R}}_{xx}\mathbf{t}_k & \mathbf{t}_h^T\hat{\mathbf{R}}_{xx}\mathbf{t}_h \end{bmatrix} \tag{4.22}$$

and

$$\mathbf{q}_i \approx \frac{1}{\hat{g}^2}\begin{bmatrix} \mathbf{t}_k^T \hat{\mathbf{R}}_{xx} \mathbf{t}_j \\ \mathbf{t}_h^T \hat{\mathbf{R}}_{xx} \mathbf{t}_j \end{bmatrix} \tag{4.23}$$

Equation (4.21), (4.22) and (4.23) indicate that c_{n1} and c_{n2} are determined by the short-time autocorrelation matrix of the input speech and the basis vectors of the DCT matrix **T**.

The decorrelating effect of the adaptive transform domain filter can be seen in Fig 4.13, in which the dotted line is the input of the adaptive transform domain filter and the solid line is its residue. Some of the peaks in the dotted line are eliminated or decreased in the solid line. The residue is the signal which can not be linearly predicted from the past ranked coefficients by the 2nd-order adaptive transform domain filter.

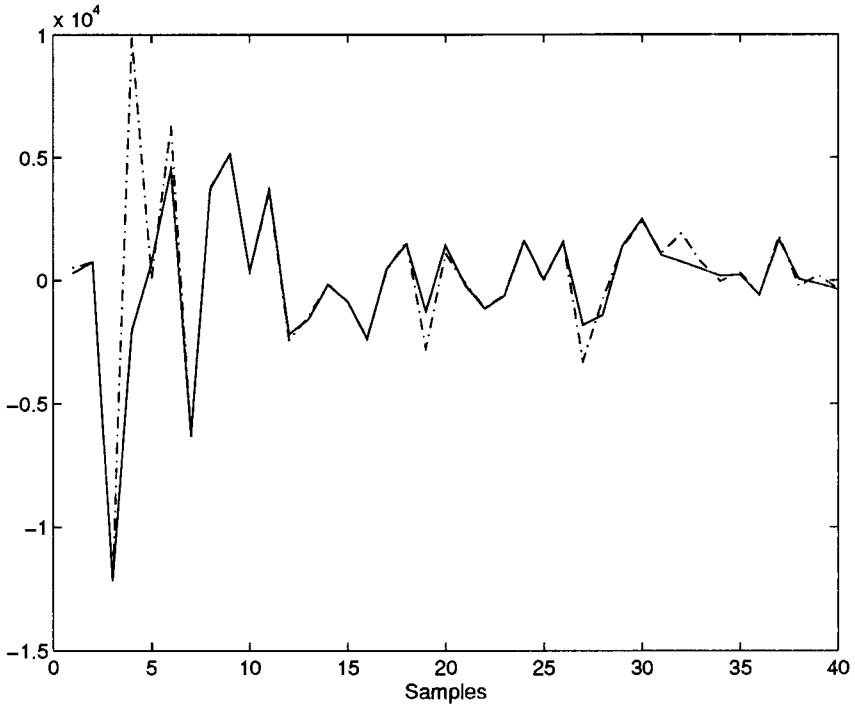


Fig 4.13 Input and output of the adaptive transform domain filter (40 points)
(Solid line: residue of the adaptive filter. Dotted line: ranked transform coefficients)

4.2.3 Bit Allocation and MSVQ

The bit allocation in the VTQ-CR-ALP is exactly the same as in the VTQ-CR. But, because of the application of the analysis-by-synthesis method, for the voiced subframe, the two-stage VQ is used for the quantization of the excitation of the synthesis filter, instead of the quantization of the truncated transform vectors as in the VTQ-CR. The structure for the two-stage VQ is depicted in Fig 4.14. The closed-loop codebook search is used in each stage to find an optimal excitation signal. The excitation vector in the first stage, $\mathbf{r}_1$, is selected such that the distortion ε_1 between the truncated transform vector $\mathbf{y}$ and the corresponding output, $\hat{\mathbf{y}}_1$, of the filter $1/C(n,z)$ is minimized.

$$\varepsilon_1 = \|\mathbf{y} - \hat{\mathbf{y}}_1\|^2 \tag{4.24}$$

In the second stage, the excitation vector, $\mathbf{r}_2$, is selected such that the distortion ε_2 between the vector $\mathbf{y} - \hat{\mathbf{y}}_1$ and the corresponding output, $\hat{\mathbf{y}}_2$, of the filter $1/C(n,z)$ is minimized.

$$\varepsilon_2 = \varepsilon_y = \|\mathbf{y} - \hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2\|^2 \tag{4.25}$$

The selected excitation vector $\mathbf{r}$ should be the sum of $\mathbf{r}_1$ and $\mathbf{r}_2$, i.e.,

$$\mathbf{r} = \mathbf{r}_1 + \mathbf{r}_2 \tag{4.26}$$

The corresponding reproduction is $\hat{\mathbf{y}} = \hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2$. The overall quantization error $\mathbf{y} - \hat{\mathbf{y}}$ is equal to the quantization residual from the second stage.

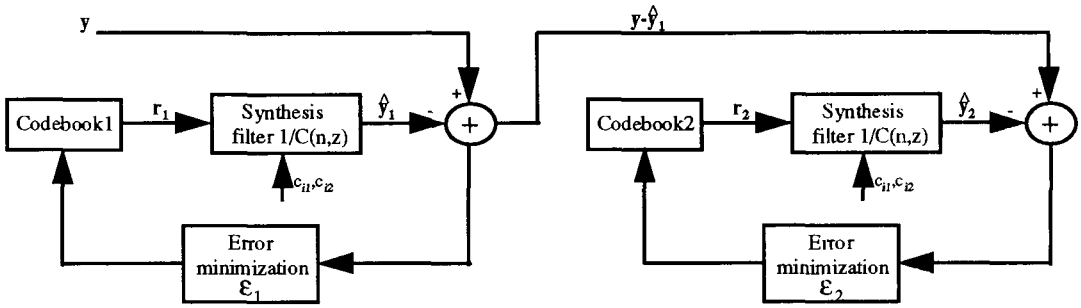


Fig 4.14 Two-stage VQ in VTQ-CR-ALP

The codebooks are trained sequentially. The training data for the first stage VQ codebook consist of the output vectors of the filter $C(n,z)$. The second stage uses a training sequence consisting of the error vectors between the residue vectors of the filter $C(n,z)$ and the corresponding codevectors selected by the first stage closed-loop codebook search. This system has higher encoding complexity compared with the VTQ-CR system, due to the closed-loop codebook search procedure.

Chapter 5

Simulations

This chapter presents simulation results for a Gauss-Markov source and for real speech waveforms. The simulations are implemented in the C language and run on a SUN workstation. Section 5.1 describes the performance criterion. Section 5.2 presents the simulation results for a Gauss-Markov process and Section 5.3 presents the simulation results for real speech waveforms.

5.1 Performance Criterion

In evaluating the quality of a speech coding system, it is important to obtain an objective measure that correctly represents the quality as perceived by the human ear. The criterion used in this thesis is the signal-to-noise ratio (SNR). It is defined as the ratio of the energy

of the input signal and the energy of the corresponding error signal. If $x(n)$ denotes the sampled input speech and $e(n)$ denotes the error between $x(n)$ and the reconstructed signal $\hat{x}(n)$, the SNR is defined as

$$SNR = 10\log_{10}\frac{\sigma_x^2}{\sigma_e^2} \tag{5.1}$$

where σ_x^2 and σ_e^2 are the variances of $x(n)$ and $e(n)$, respectively. This SNR criterion is mainly influenced by high-level segments of speech, i.e., by voiced sounds, and it reflects only to a small degree the coder performance for unvoiced sounds and for the idle channel situation.

The system coding gain is introduced here to evaluate the compression performance of the coding systems. The average MSE distortion D of the coding scheme is defined as

$$D = \frac{1}{M}E\left[\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|^2\right] \tag{5.2}$$

Let D_{PCM} be the distortion of the Pulse-Code Modulation (PCM) scheme and D_{VTQ} be the distortion of the VTQ system. The VTQ coding gain over the PCM is defined by [23][27]

$$G_{VTQ} = \frac{D_{PCM}}{D_{VTQ}} \tag{5.3}$$

Similarly, the VTQ-CR and VTQ-CR-ALP system coding gains over the PCM can be defined as

$$G_{VTQ-CR} = \frac{D_{PCM}}{D_{VTQ-CR}} \tag{5.4}$$

$$G_{VTQ-CR-ALP} = \frac{D_{PCM}}{D_{VTQ-CR-ALP}} \tag{5.5}$$

where D_{VTQ-CR} and $D_{VTQ-CR-ALP}$ are the corresponding distortions in the VTQ-CR and VTQ-CR-ALP, respectively. From Equation (5.3)--(5.5), the VTQ-CR and VTQ-CR-ALP coding gains over VTQ can be obtained, respectively, as

$$G^{*}_{VTQ-CR} = \frac{G_{VTQ-CR}}{G_{VTQ}} = \frac{D_{VTQ}}{D_{VTQ-CR}} \tag{5.6}$$

$$G^{*}_{VTQ-CR-ALP} = \frac{G_{VTQ-CR-ALP}}{G_{VTQ}} = \frac{D_{VTQ}}{D_{VTQ-CR-ALP}} \tag{5.7}$$

In terms of SNR measure, G^{*}_{VTQ-CR} and $G^{*}_{VTQ-CR-ALP}$ can be computed by

$$G^{*}_{VTQ-CR}(dB) = SNR_{VTQ-CR} - SNR_{VTQ} \tag{5.8}$$

$$G^{*}_{VTQ-CR-ALP}(dB) = SNR_{VTQ-CR-ALP} - SNR_{VTQ} \tag{5.9}$$

Equation (5.8) and (5.9) indicate that the coding gains over VTQ measure the SNR increases due to the adoption of the ranking VQ and the adaptive transform domain linear prediction.

Under the assumption that the transform coefficients are independent, it can be proven [27] that

$$G_{VTQ} = G_{TC} \cdot G_{VQ} \tag{5.10}$$

where G_{TC} is the transform coding gain with scalar quantization and G_{VQ} is the gain due to the use of vector quantizers instead of scalar quantizers. G_{TC} is equal to the ratio of the arithmetic and the geometric mean of the variances of the transform coefficients [23].

$$G_{TC} = \frac{\frac{1}{M}\sum_{i=1}^{M}\sigma_i^2}{\left[\prod_{i=1}^{M}\sigma_i^2\right]^{1/M}} \tag{5.11}$$

where σ_i^2 is the variance of the *i*th transform coefficient. When Equation (5.11) is applied to measure the transform coding gain $\tilde{G}_{TC}$ for ranked VTQ, σ_i^2 should be the variance of the *i*th ranked transform coefficient. The ranked VTQ transform coding gain over VTQ is obtained by

$$G_{TC}^{*}(dB) = \tilde{G}_{TC}(dB) - G_{TC}(dB) \tag{5.12}$$

G_{VQ} is the VQ gain which depends on the type of the chosen quantizer. For the VTQ system in which the split VQ with the optimal bit allocation is used, the expression of the G_{VQ} was derived by Cuperman in [27]. It only depends on the bit allocation and the pdf of the transform coefficients. For the VTQ-CR and VTQ-CR-ALP systems, it is difficult to derive the mathematical expressions of the G_{VQ} due to the truncation of the transform coefficients and the applications of the multistage VQ and the adaptive transform domain linear prediction. The G_{VQ} can be estimated through system coding gain and transform coding gain.

5.2 Simulation on the Gauss-Markov Source

The purpose of simulation on the Gauss-Markov source is to test the effectiveness of the coefficient ranking model in the transform coding systems. First, we need to build a Gauss-Markov model which generates data sequences used as the input signal of the systems. Second, the simulations are performed to compare the performance of the VQ of the ranked transform coefficients with the performance of the VQ of the non-ranked transform coefficients.

5.2.1 The Design of a Gauss-Markov Model

We build a Gauss-Markov model which generates data sequences having spectral characteristics of voiced speech. Fig 5.1 is the block diagram of this model.

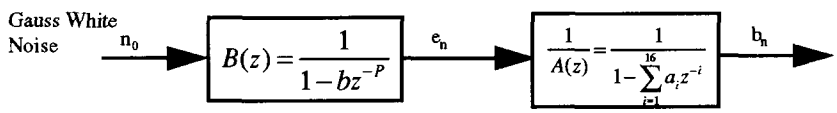


Fig 5.1 Gauss-Markov model

It is known that if the input speech of a linear prediction filter is voiced, the residual waveform will contain periodic spikes at the pitch frequency. In this model, the sequence e_n, which has the periodicity of the voiced speech, is obtained by filtering the Gauss white noise n_0 through the pitch filter $B(z)$. The parameter P controls the pitch frequency. Here, we set $P=81$ and the gain of the filter $b=0.9$. The 16th-order linear prediction coefficients $a_i, i=1,2,...16$, are calculated from several frames of typical voiced speech

signal. Fig 5.2 shows the frequency spectrum of $1/A(z)$. Fig 5.3 is a segment of data sequence b_n generated by such model.

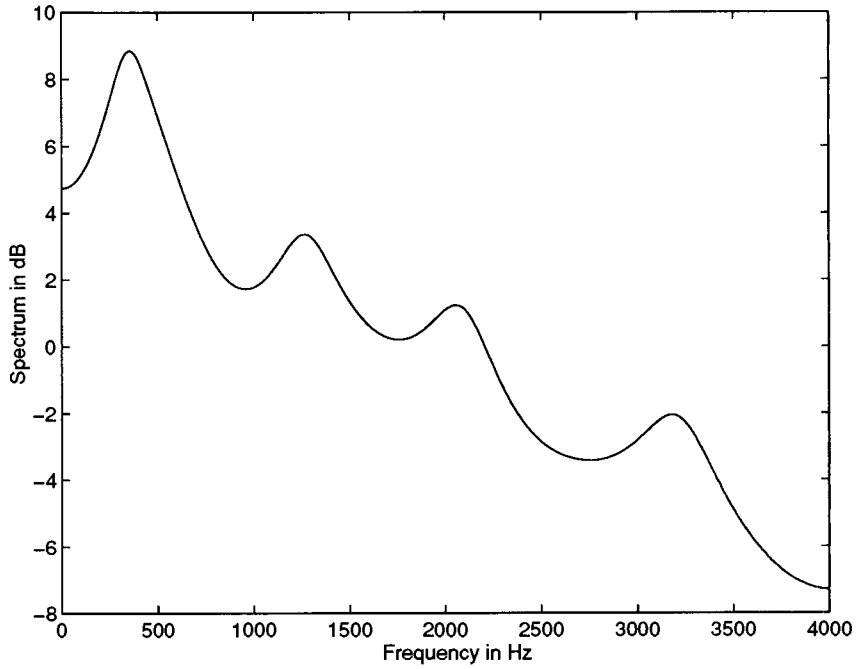


Fig 5.2 The frequency spectrum of $1/A(z)$

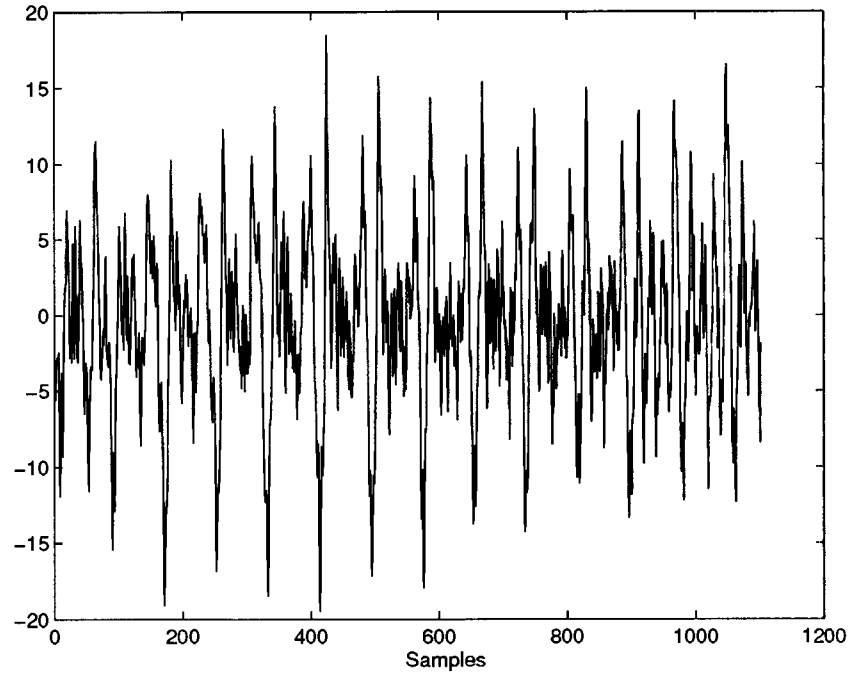


Fig 5.3 Data sequence of the Gauss-Markov source

A sequence of 3,900,000 samples is used as the training data. A separate sequence of 200,000 is used for testing purpose.

5.2.2 Simulation Results on the Gauss-Markov Source

In the simulation with the VTQ system at rate 2.4 kbps, each subframe of transform coefficients have dimension $M = 128$. Bit allocation is optimized by the procedure described in Chapter 3. Each transform vector is divided into 4 sub-vectors, i.e., $m = 4$, for vector quantization, having dimension $k_1 = 7$, $k_2 = 8$, $k_3 = 42$, $k_4 = 71$. With 36 available bits for the quantization of each subframe, the optimal bit assignment is $B_1 = 10$, $B_2 = 10$, $B_3 = 10$, $B_4 = 6$. The limitation on the codebook size is 1024.

For the simulation with the VTQ-CR system, the transform coefficients are ranked before VQ. We found that the first 40 out of the 128 ranked coefficients concentrate the 94.1% of the signal energy. Therefore, the ranked transform vectors are truncated and only the first 40 data are vector quantized and the rest is set to be zero. Since all input frames are "voiced", there is no voicing decision to be made. Also, the linear prediction coefficients and the pitch period are constant, therefore, it is not necessary to transmit them to the decoder. Here, the available bits for quantizing each truncated transform vector is 36, too. A four-stage VQ is designed with 9 bits assigned to each stage and the dimension is 40. The quantization is implemented in four successive stages, in which the quantization error from the previous stage is used as input to the next stage of the VQ. The identical gain codebook is used in both systems and 5 bits are assigned for the scalar quantization of the gain.

The SNR performances for different data sequence are presented in Table 5.1. Test 1 and test 2 are the results obtained for the training data and test 3 to test 5 are the results for the testing sequence.

	Test 1	Test 2	Test 3	Test 4	Test 5	Average
SNR(dB) (Non-ranked)	11.23	11.10	10.78	10.54	10.21	10.772
SNR(dB) (Ranked)	11.82	11.79	11.57	11.41	10.72	11.462

Table 5.1 The simulation results on the Gauss-Markov source

Compared with the non-ranked VTQ system, the VQ of the ranked transform coefficients could provide average 0.7 dB SNR improvement, which gives a good indication about the efficiency of the coefficient order ranking approach. It should be pointed out that the simulation results on the Gauss-Markov source make only a limited sense on the evaluation of the VTQ and the VTQ-CR systems, because the input waveform unlike the speech signal has no spectral changes and, hence, the ranking is fixed. However, we still can conclude that ranking transform coefficients in a descending order of their energy values and vector quantizing the most significant coefficients can make the VQ of the transform coefficients more efficient at a low bit rate.

Theoretically, VTQ and ranked VTQ transform coding gains over PCM are the same, i.e., $G_{TC}^{*} = 0\ dB$, if the input signal is stationary. However, it can be noticed that there exists slight spectrum changes when the autocorrelation functions of the input signal are estimated for each frame. The VTQ and ranked VTQ transform coding gains over PCM are 5.78 dB and 6.05 dB, respectively, which are computed using the estimated variances $\hat{\sigma}_i^2$ in Equation (5.11). Therefore, the theoretical value of the system coding gain is 0.27

dB if the VQ gain in Equation (5.10) is neglected. The experimental value of the coding gain is approximately 0.7 dB, which is obtained by using Equation (5.8). These results indicate that the ranked VTQ can provide a slightly higher transform coding gain over the non-ranked VTQ. However, the gain is small because of the fact that the signal is stationary.

Fig 5.5 shows the transform coefficients of the input waveform shown in Fig 5.4. Fig 5.6 is its ranked sequence. Fig 5.7 and 5.8 show the corresponding reconstructed signals for the VTQ system and the VTQ-CR system, respectively.

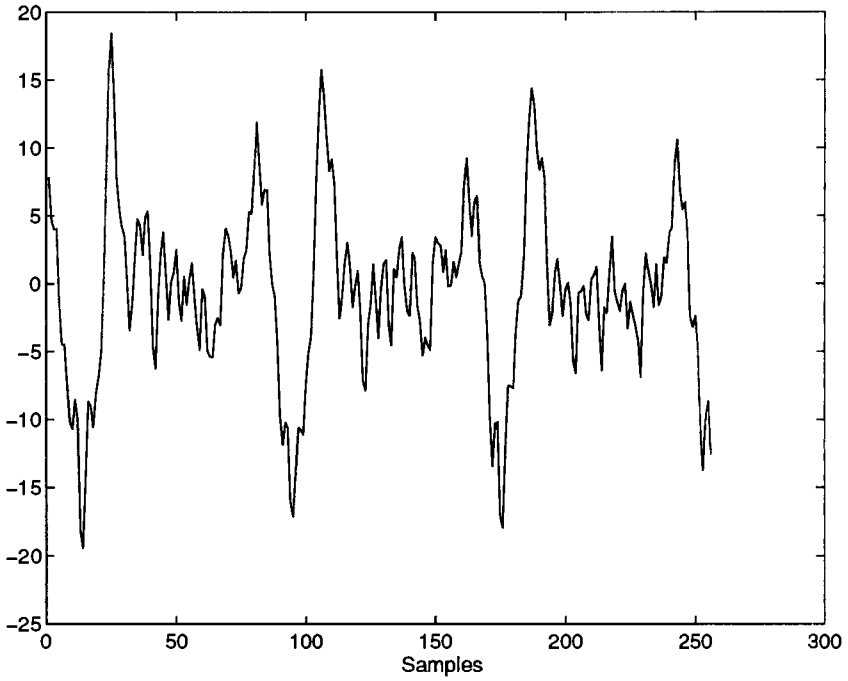


Fig 5.4 Waveform of the Gauss-Markov source

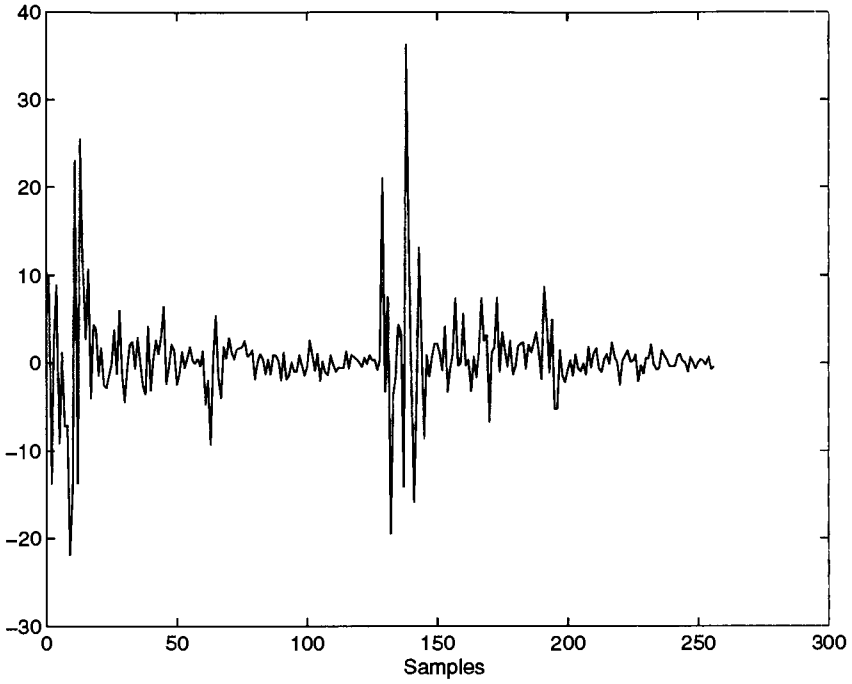


Fig 5.5 Transform coefficients of the Gauss-Markov source

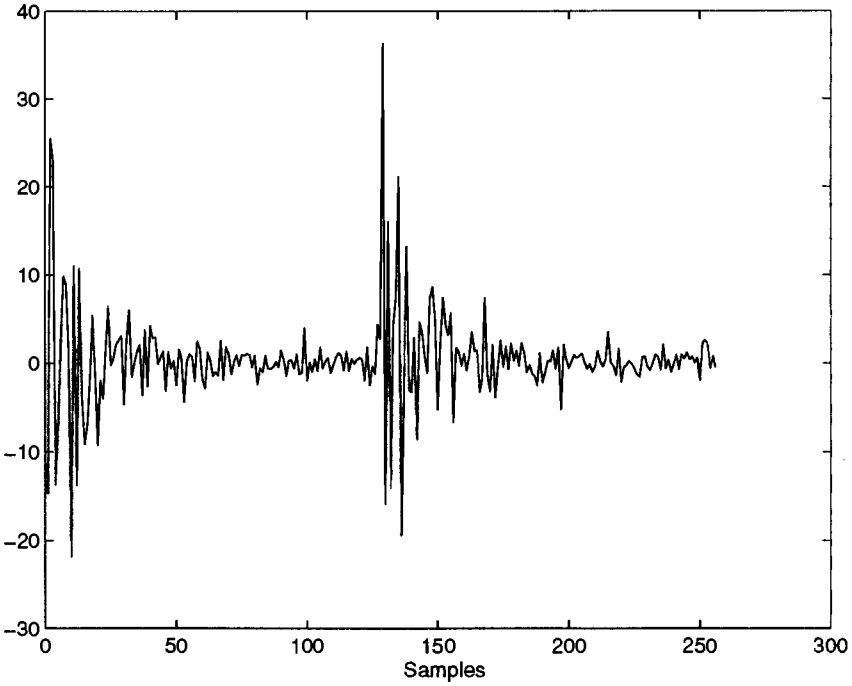


Fig 5.6 Ranked transform coefficients of the Gauss-Markov source

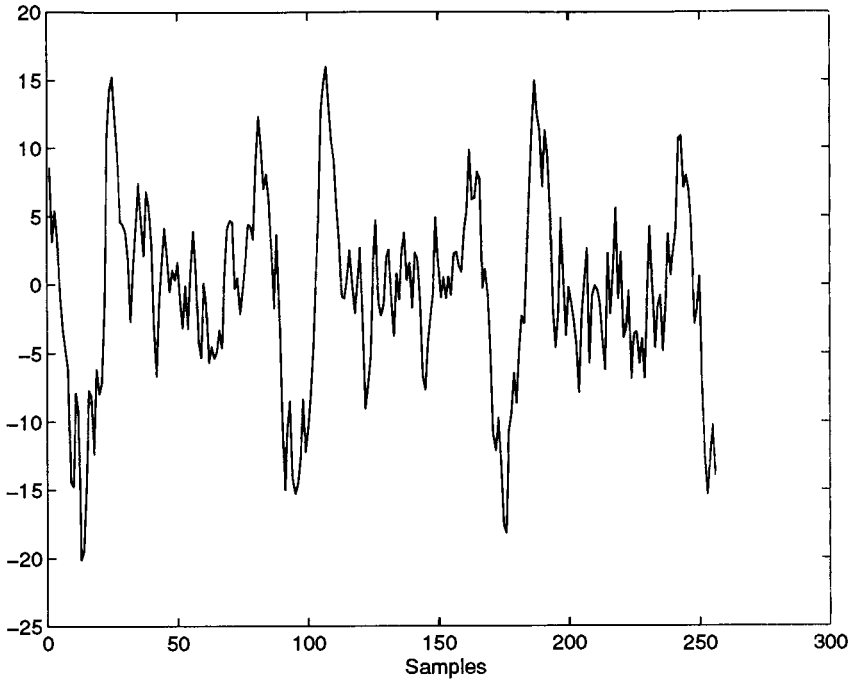


Fig 5.7 Reproduced waveform of the Gauss-Markov source (VTQ)

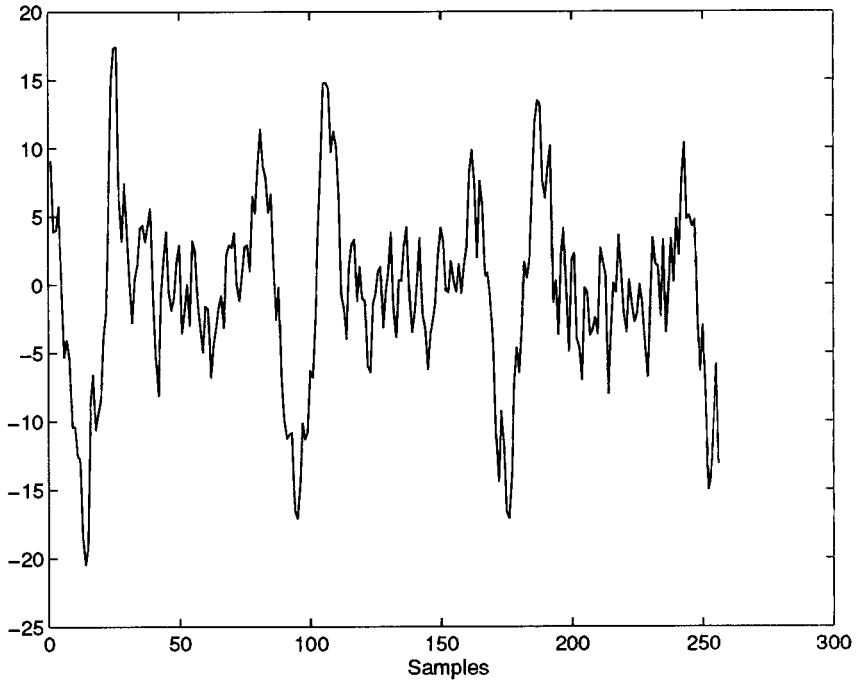


Fig 5.8 Reproduced waveform of the Gauss-Markov source (VTQ-CR)

When the input signal is the real speech waveform, the SNR performance of the VTQ system will degrade since the non-stationariness of the speech signal is not taken into account in the estimation of the variances of the transform coefficients in the optimal bit allocation -- this problem does not exist in the simulation on the Gauss-Markov source. On the other hand, in the case of real speech input, some bit codes have to be used for the transmission of the speech spectral information in the VTQ-CR system -- this transmission does not needed in the simulation on the Gauss-Markov source. We will give the simulation results on the real speech source in Section 5.3.

5.3 Simulation on the Real Speech Source

The simulation purpose on the real speech source is to evaluate the performance of the speech coding systems, VTQ, VTQ-CR, and VTQ-CR-ALP. A file containing 26015×128 speech samples from different talkers is used for the VQ codebook training. Different speech content from the same talkers and material from talkers not included in training data is used to test the system performance.

The input speech signal sampled at 8 kHz is segmented in frames of 256 samples. Each frame is divided into two subframes with each having 128 samples. At the bit rate of 2.4 kbps, the 256 samples in a frame are compressed and represented by 77 bits. For the VQ of each subframe transform coefficients, 4 split VQ is applied and 36 bits are available. 5

bits are used for the scalar quantization of the gain. The maximal codebook size is set to 1024. The optimal bit allocation for the VTQ system is derived as follows:

m	1	2	3	4
Dimension (k_i)	27	23	33	45
Bit Assigned (B_i)	10	10	8	8

Table 5.2 Optimal bit allocation for the VQ of the transform coefficients in VTQ

In the simulation with the VTQ-CR and VTQ-CR-ALP systems, the transform coefficient truncation is based on the estimated vector variances or energy distributions. For a voiced subframe, statistically, the first 50 out of the 128 ranked transform coefficients concentrate 91.2% of the subframe energy. Therefore, these 50 coefficients are vector quantized and the rest of 78 coefficients is set to zero. According to the bit allocation strategy described in Chapter 4, 20 bits are available for the VQ of the ranked coefficients. A two-stage vector quantizer is designed with 11 bits in the first stage and 9 bits in the second stage. Table 5.3 shows the detailed bit allocation for VTQ-CR and VTQ-CR-ALP.

	Voiced (bits)	Unvoiced (bits)
LSP	24	24
U/V	1	1
Pitch	7	0
Gain	5	7
Transform Vector or Excitation	1st stage: 2*11 2nd stage: 2*9	1st subvector:2*11 2nd subvector:2*11

Table 5.3 The detailed bit allocation for VTQ and VTQ-CR-ALP

The bit assignment is subjected to the constraints of the complexity, the SNR performance and the listening subjective quality of the coded speech. In the VTQ-CR-ALP system, because of the application of the analysis-by-synthesis method, the two-stage VQ is used for the quantization of the excitation of the synthesis filter, instead of the quantization of the truncated transform vectors as in the VTQ-CR system. Closed-loop codebook search is adopted to find the optimal excitation for the synthesis filter.

For the unvoiced subframe, since the energy distribution of the ranked transform coefficients is quite scattered, the last 20 coefficients are set to zero and the first 108 are vector quantized. We devide these 108 coefficients into two sub-vectors with one having the first 44 and another having the rest of 64. With 22 bits available as discussed in Chapter 4, split VQ is used with each sub-vector 11 bits assigned. The identical unvoiced codebooks are used in both systems.

5 sets of experiments are made. SNRs of the three systems are recorded in Table 5.4. Test 1 and test 2 are the results obtained inside of the training process and test 3 to test 5 are from outside of the training sequence.

	Test 1	Test 2	Test 3	Test 4	Test 5	Average
SNR(dB) (VTQ)	5.36	5.12	5.05	2.96	2.32	4.162
SNR(dB) (VTQ-CR.)	6.57	6.49	5.85	3.88	3.81	5.320
SNR(dB) (VTQ-CR-ALP)	7.47	7.26	6.62	4.79	4.45	6.118

Table 5.4 The simulation results on real speech source

The simulation results indicate a performance improvement of 1-1.5 dB for the transform coefficient order ranking in VTQ-CR when compared with coefficient non-ranking scheme in VTQ. For a subframe of 128 voiced ranked transform coefficients, the energy concentrated in the first 50 components is 91.2% of the subframe energy, while the first 50 non-ranked voiced coefficients just have approximately 60% of the energy. At low bit rates, coding the most significant coefficients which are in the descending order of their energy values can make the VQ more efficient. VTQ-CR and VTQ-CR-ALP outperform VTQ, although they use less bits than VTQ to quantize the coefficient sequence because of the transmission of the LSPs, gain, voicing decision and pitch period.

From Table 5.4, we can see that VTQ-CR-ALP improves the SNR performance of the VTQ-CR system further by 0.6-0.9 dB. An adaptive transform domain linear predictor decorrelates the voiced ranked transform coefficients. The VTQ-CR-ALP system combines the features of both coefficient order ranking and adaptive transform domain linear prediction to form a more sophisticated coding scheme which provides better quality and more efficient speech coding. MSVQ coupled with codebook closed-loop search is used to obtain an efficient, high quality and low complexity quantizer.

For the real speech source, the VTQ and VTQ-CR transform coding gains over PCM are 4.17 dB and 4.79 dB, respectively, which are computed using the estimated variances $\hat{\sigma}_i^2$ in Equation (5.11). Therefore, the theoretical value of the VTQ-CR and the VTQ-CR-ALP system coding gains over VTQ are 0.62 dB if the VQ gains in Equation (5.10) are neglected. The experimental values of the coding gains are 1.158 dB and 1.956 dB, respectively, which are obtained by using Equation (5.8) and (5.9). Compared with the

results obtained by the Gauss-Markov source, we can see that the ranked VQ is more efficient for the non-stationary source.

When the input signal only consists of voiced speech, the corresponding transform coding gains over PCM are increased to 5.63 dB for VTQ and 6.44 dB for VTQ-CR. These results indicate that unvoiced speech decreases the VTQ and VTQ-CR transform coding gains. Since unvoiced speech is more like random "noise", the transform coding and the coefficient ranking is less efficient.

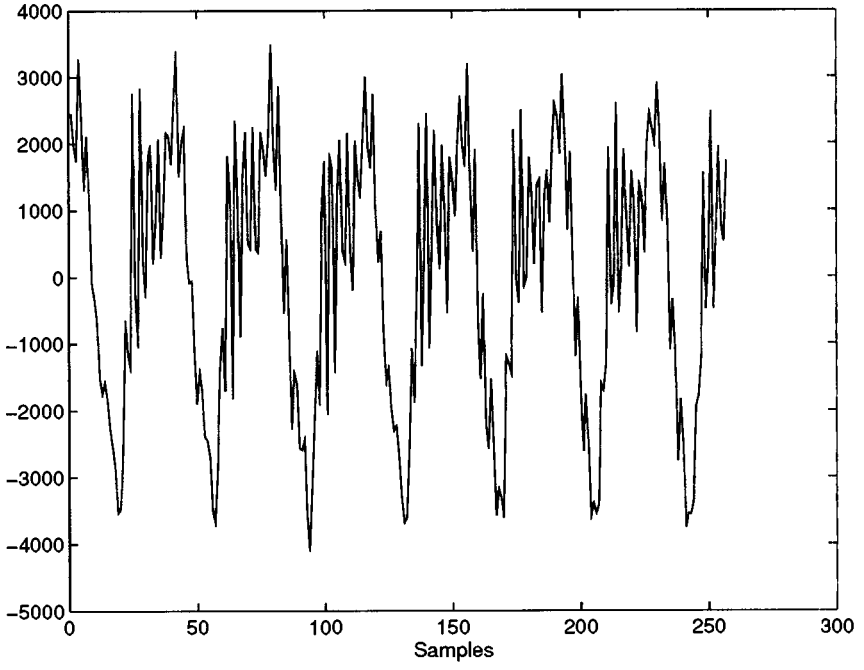


Fig 5.9 Voiced speech signal

Fig 5.9 shows two subframes of voiced speech signal. Fig 5.10 to Fig 5.12 show the corresponding reproduced waveforms generated by VTQ, VTQ-CR and VTQ-CR-ALP, respectively. Fig 5.13 shows a frame of unvoiced speech. The corresponding reproduced waveforms generated by VTQ and VTQ-CR are shown in the Fig 5.14 and Fig 5.15,

respectively. Fig 5.16 shows the SEGSNR variation with speech subframes in VTQ-CR-ALP. Here, the SEGSNR is evaluated by computing the SNR for each subframe speech.

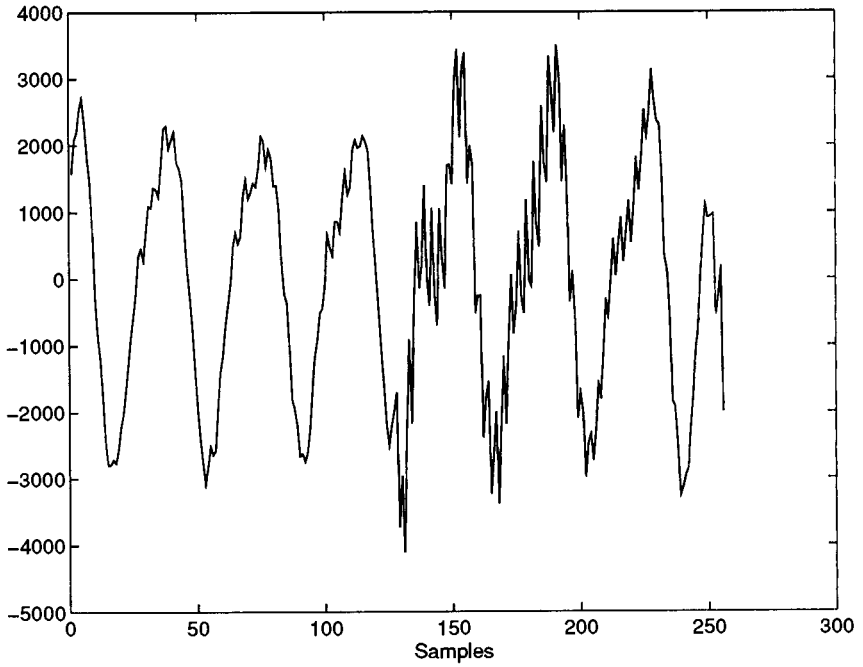


Fig 5.10 Reproduced voiced speech signal (VTQ)

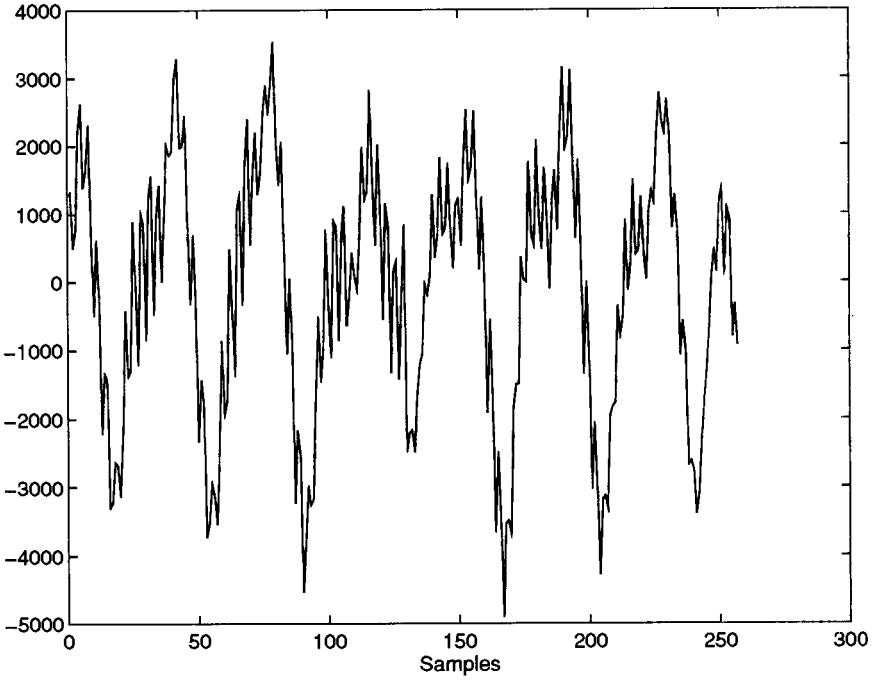


Fig 5.11 Reproduced voiced speech signal (VTQ-CR)

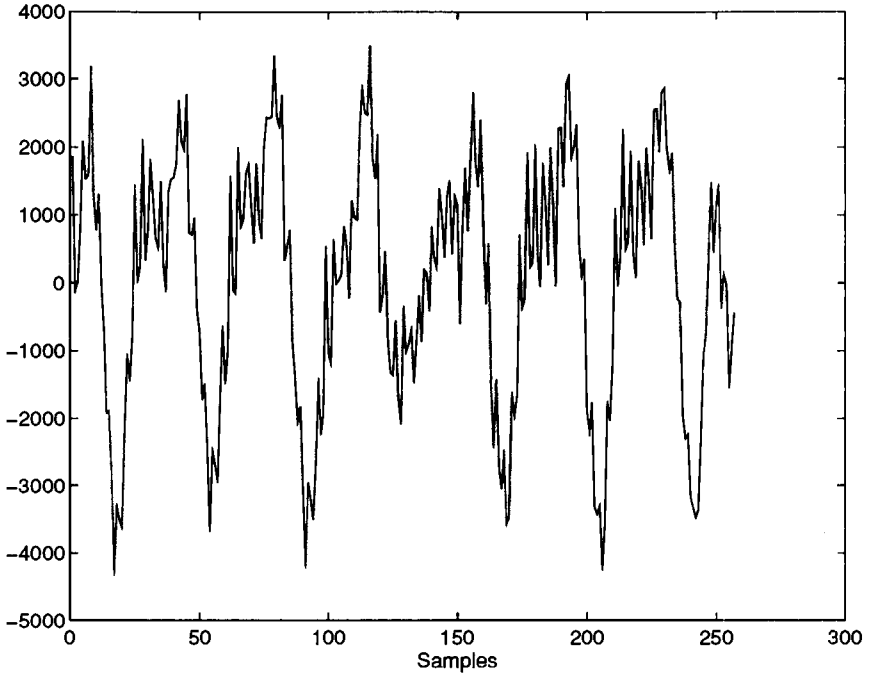


Fig 5.12 Reproduced voiced speech signal (VTQ-CR-ALP)

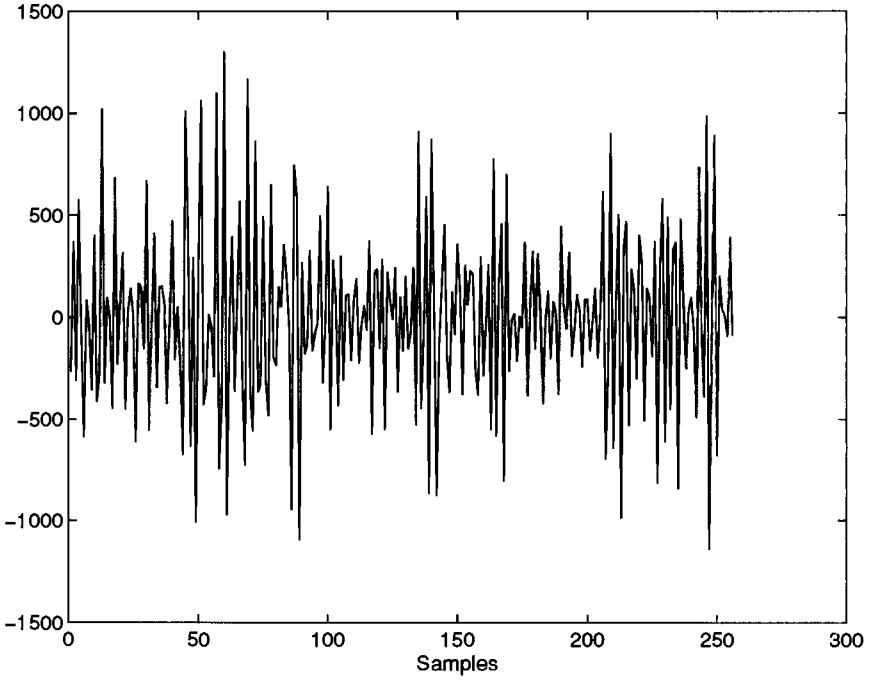


Fig 5.13 Unvoiced speech signal

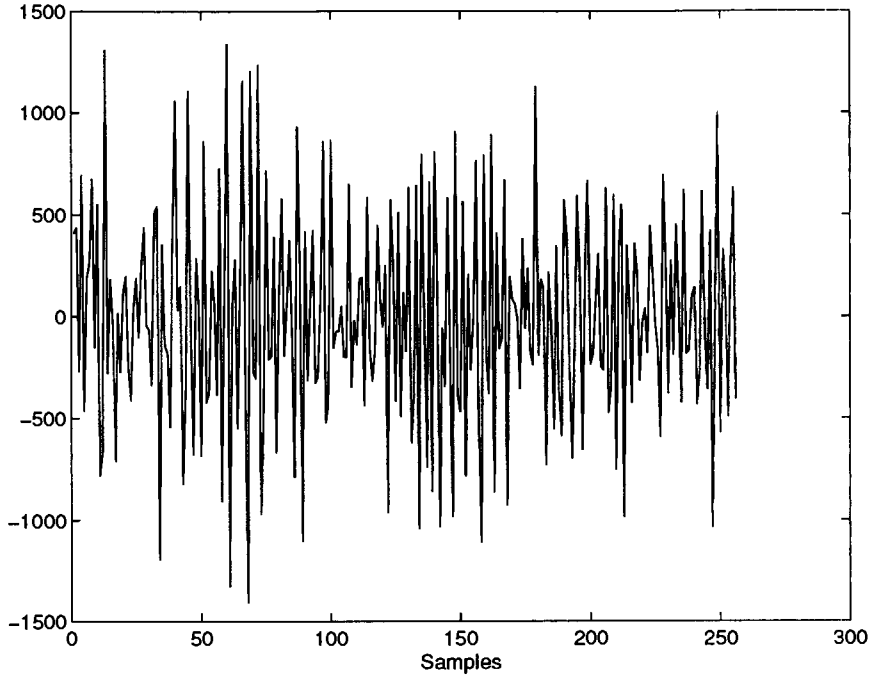


Fig 5.14 Reproduced unvoiced speech signal (VTQ)

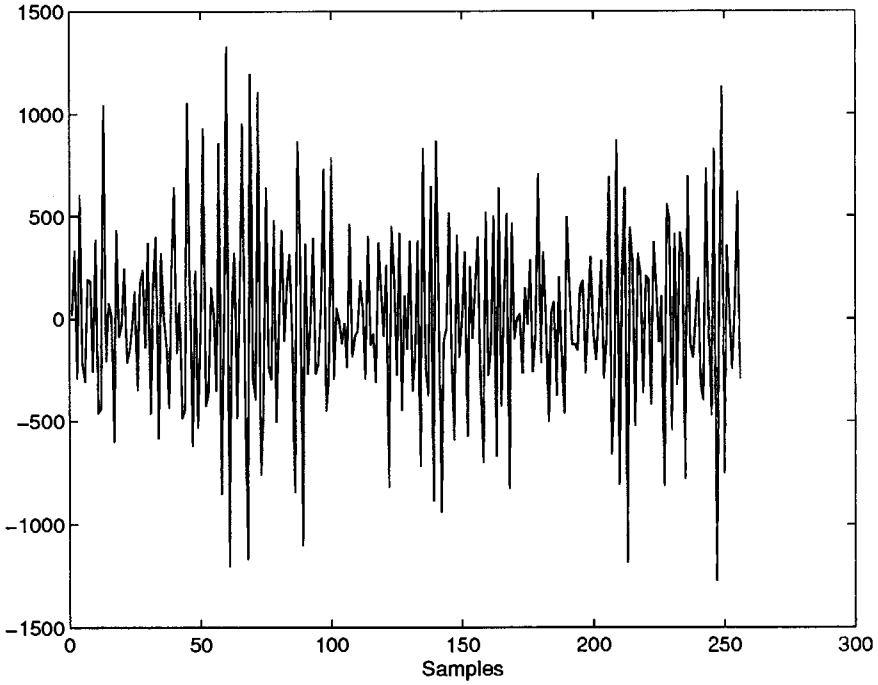


Fig 5.15 Reproduced unvoiced speech signal (VTQ-CR)

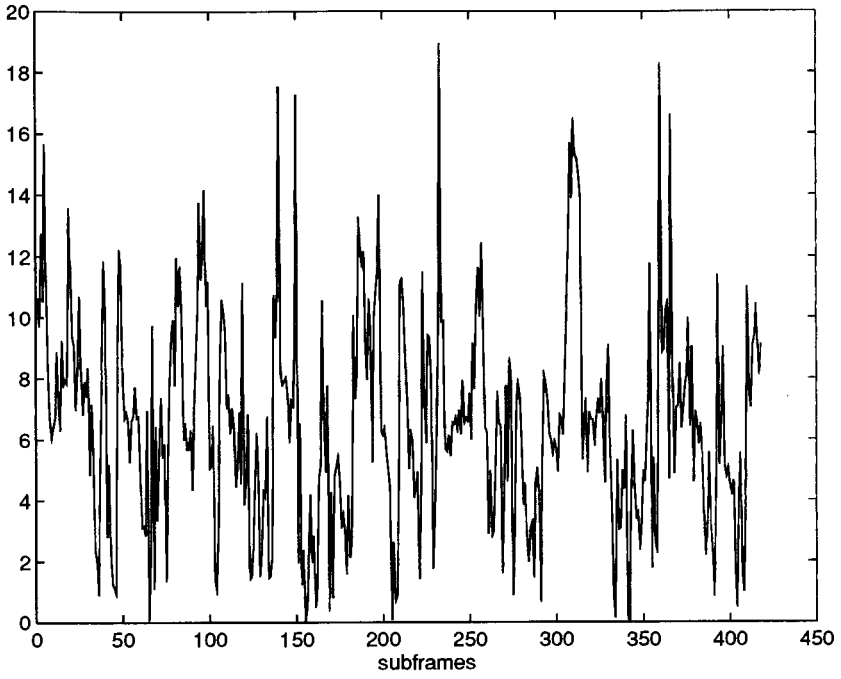


Fig 5.16 SEGSNR variation with speech subframes (VTQ-CR-ALP)

Chapter 6

Conclusions

In this thesis, two transform coding systems based on the coefficient ranking model are developed: VTQ-CR and VTQ-CR-ALP. Coefficient ranking technique and adaptive linear prediction analysis are proposed to improve the performance of conventional VTQ coders. Particular emphasis is placed on the effectiveness and efficiency of the ranking vector quantization and the application of the adaptive transform domain linear predictor in the transform coding at low bit rate. A lot of work has been done on the system designs and the performance analysis.

By comparing the performance of the coefficient ranking VTQ for a Gauss-Markov source with the performance of a conventional VTQ system, it is shown that ranking transform coefficients in a descending order of their energy values and vector quantizing the most significant coefficients can make the VQ more efficient at low bit rate. The comparison of the performances of these three systems for real speech signal indicates that

VTQ-CR and VTQ-CR-ALP outperform VTQ, although they use less bits than VTQ to quantize the coefficient sequence because of the transmission of the LSPs, voicing decision and pitch period. In the low bit transform coding systems, the VTQ coding scheme could not quantize the transform coefficients effectively because the non-stationariness of the speech signal is not taken into account in the estimation of the variances of the transform coefficients in the optimal bit allocation.

A further performance improvement can be achieved by applying an adaptive transform domain linear predictor to the voiced ranked coefficients, where the correlations between the coefficients are reduced. MSVQ coupled with the closed-loop VQ codebook search is used to obtain an efficient, high quality and low complexity quantizer. VTQ-CR-ALP combines the features of both coefficient order ranking and linear prediction to form a more sophisticated coding scheme which provides better quality and more efficient speech coding.

References

[1] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates", *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, pp 614-617, May 1982.

[2] B. S. Atal and M. R. Schroeder, "Stochastic coding of speech signals at very low bit rates", *Proc. Int. Conf. Commun.*, vol. ICC84, part 2, pp 1610-1613, May 1984.

[3] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): high quality speech at very low bit rates", *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, pp 937-940, March 1985.

[4] P. Kroon, E. F. Deprettere and R. J. Sluyter, "Regular-pulse excitation: a novel approach to effective and efficient multi-pulse coding of speech", *IEEE Trans. ASSP*, vol. ASSP-34, pp 1054-1063 (1986).

[5] G. Davidson and A. Gersho, "Complexity reduction method for vector excitation coding", *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 4, pp 3055-3058, April 1986.

[6] P. Kroon and E. F. Deprettere, "A class of analysis-by-synthesis predictive coders for high quality speech coding at rates between 4.8 and 16 kbit/s", *IEEE Jour. on Selected Areas in Commun.*, vol. 6, no. 2, pp 353-363, Feb. 1988.

[7] D. P. Kemp, R. A. Sueda, and T. E. Tremain, "An evaluation of 4800 bps voice coders", *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, pp 200-203, March 1989.

[8] J. P. Campbell, Jr., V. C. Welch and T. E. Tremain, "An expandable error-protected 4800 bps CELP coder (U.S. Federal Standard 4800 bps voice coder)", *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 2, pp 735-738, March 1989.

[9] K. Ozawa and T. Araseki, "High quality multi-pulse speech coder with pitch prediction", *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, pp 1689-1692, April 1986.

[10] S. Singhal and B. S. Atal, "Amplitude optimization and pitch prediction in multipulse coders", *IEEE Trans. ASSP*, vol. 37, pp 317-327 (1989).

[11] A. Gersho, "Advances in speech and audio compression", *Proc. IEEE*, vol. 82, pp 900-918, 1994.

[12] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, New York, NY: Springer-Verlag, 1976.

[13] T. E. Tremain, "The Government Standard Linear Predictive Coding Algorithm: LPC-10", *Speech Technology*, pp 40-49, April 1982.

[14] "Telecommunications: Analog to digital conversion of voice by 2400 bit/sec. Linear Predictive Coding, FED-STD-1015", *Office of Technology and Standards, National Communications System,* Washington DC (March 1983).

[15] A. V. McCree and T. P. Barnwell, "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding", *IEEE Trans. on Speech and Audio Processing,* vol. 3, No. 4, pp 242-249, July 1995.

[16] A. V. McCree and T. P. Barnwell, "Implementation and Evaluation of a 2400 bps Mixed Excitation LPC Vocoder", *Proc. Int. Conf. on Acoustics, Speech and Signal Processing,* pp II-159-II-162, April 1993.

[17] Telecommunications: Analog to Digital Conversion of Radio Voice by 4800 bit/s: Code Excited Linear Prediction (CELP) FED-STD-1016, *Office of Technical Standards, National Communications System,* Washington DC, Nov. 1989.

[18] Coding of Speech at 16 kbit/s Using Low-Delay Code Excited Linear Prediction, ITU-T Recommendation G. 728, Geneva, Switzland, Sept. 1992.

[19] Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP), ITU-T Recommendation G. 729, Geneva, Switzland, 1995.

[20] B. S. Atal and B. E. Caspers, "Beyond multipulse and CELP towards high quality speech at 4 kb/s", in *Advance in Speech Coding*, Ed. by B. S. Atal, V. Cuperman and A. Gersho, Klumer Academic Publishers, Massachusetts, pp 191-201, 1991.

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

[22] C. Rowden, *Speech Processing,* McGraw-Hill Book, 1992

[23] R. Zelinski and P. Noll: 'Adaptive transform coding of speech signals', *IEEE Trans Acoustics, Speech and Signal Processing*, ASSP--25, No 4, pp 299-309 August 1977.

[24] Keith A. Teague, Bryce Leach and Walter Andrews, "Development of a High-Quality MBE Based Vocoder for Implementation 2400 bps", *Proc. of the IEEE Wichita Conf. on Commun. Networking, and Signal Processing,* April, 1994.

[25] Y. Shoham, "Speech Coding at 2.4 kbps and Below Via Time-Frequency Interpolation", *Proc. IEEE Workshop on Speech Coding for Telecommun.,* pp107-108, Oct. 1993.

[26] R. J. McAulay and T. F. Quatieri, "The Application of Subband Coding to Improve Quality and Robustness of the Sinusoidal Transform Coder", *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing*, ppII-439--II-446, April 1993.

[27] V. Cuperman, "On Adaptive Vector Transform Quantization for Speech Coding", *IEEE Trans. on Commun.,* vol. 37, No.3, March 1989.

[28] A. Gersho and R. Gray, *Vector Quantization and Signal Compression,* Kluwer Academic Publishers, 1992.

[29] Y. Cheng, A. Gersho, B. Ramamurthi and Y. Shoham, "Fast search algorithms for vector quantization and pattern matching", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (San Diego, CA, Mar. 1984), paper 9.11.

[30] A. Gersho, "On the structure of vector quantizers", *IEEE trans. Inform. Theory, vol. IT-28,* pp. 157-166, Mar.1982.

[31] B. H. Juang and A. Gray, "Multiple Stage Vector Quantization for Speech Coding", *IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* May 1982.

[32]W.-Y. Chan, S. Gupta and A. Gersho, "Enhanced Multistage Vector Quantization by Joint Codebook Design," *IEEE Trans. on Commun.,* vol. 40, No. 11, pp 1693-1697, Nov. 1992.

[33] W. P. LeBlanc, B. Bhattacharya, S. A. Mahmoud, V. Cuperman, "Efficient Search and Design Procedures for Robust Multi-Stage VQ of LPC Parameters for 4 kb/s Speech Coding", *IEEE Trans. on Speech and Audio Processing,* vol. 1, No. 4, pp 1-13, Oct. 1993.

[34] V. Cuperman, "Speech Coding", *Advances in Electronics and Electron Physics,* vol. 82, 1991, pp 97-196.

[35] N. Levinson, "The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction", *J. Math. Phys.,* pp 261-278, 1947.

[36] J. Durbin, "The Fitting of Time Series Models", *Rev. Inst. Int. Statist.,* pp 233-243, 1960.

[37] J. G. Proakis and D. G. Manolakis, *Introduction To Digital Signal Processing*, Macmillan, 1988.

[38] F. Soong and B. Juang, "Line Spectrum Pair (LSP) and Speech Data Compression", *IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1983.

[39] G. Kang and L. Fransen, "Application of Line-Spectrum Pairs To Low-Bit-Rate Speech Encoders", *IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1985.

[40] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, NJ: Prentice-Hall, 1978.

[41] J. M. Tribolet and R. E. Crochiere, "Frequency domain coding of speech", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp 512-530, Aug. 1977.

[42] N. Ahmed and K. R. Rao, *Orthogonal Transforms for Digital Signal Processing*, Berlin: Springer, 1975.

[43] J. W. Cooley and J. W. Tukey, "An Algorithm for the Machine Calculation of Complex Fourier Series", *Mathematics Computation*, vol. 19, pp 297-301, 1965.

[44] N. S. Jayant, "Waveform Quantization and Coding", *IEEE Press.*, New York, 1976.

[45] V. Cuperman, "Vector Transform Quantization for Speech Coding", *Proc. ICASSP*, pp 23.3.1-23.3.5, 1986.

[46] A. Segall, "Bit allocation and encoding for vector sources", *IEEE Trans. Inform. Theory*, vol. IT-22, pp162-169, Mar. 1976.

[47] B. Bunin, "Rate-distortion functions for Gaussian-Markov processes", *Bell Tech. J.*, pp 3059-3074, Nov. 1969.

[48] Peter Kabal and Ravi Prakash Ramachandran, "The Computation of Line Spectral Frequencies Using Chebyshev Polynomials", *IEEE Trans Acoust., Speech, Signal Processing*, vol.34, no.6, Dec. 1986, pp. 1419-1426.

[49] J. Chen and A. Gersho, "Real-time vector APC speech coding at 4800 bps with adaptive postfiltering ", *Proc. ICASSP*, pp 2185-2188, 1987.

[50] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave", *J. Acoust. Soc. Amer.*, vol. 50, No. 2, pp 637-655, Aug. 1971.

[51] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies", *Electron. Commun. Japan*, vol. 53-A, pp 36-43, 1970.

[52] J. Makhoul, "Linear prediction: A tutorial review", *Proc. IEEE*, vol. 63, No. 4, pp 561-580, Apr. 1975.

[53] G. S. Kang and D. C. Coulter, "600 bps voice digitizer", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp 91-94, Philadelphia, PA, Apr. 1976.

[54] Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design", *IEEE Trans. Commun.*, vol. COM-28, No. 1, pp 84-95, Jan. 1980.

[55] A. Buzo, A. H. Gray Jr., R. M. Gray and J. D. Markel, "Speech coding based upon vector quantization ", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, No. 5, pp 562-574, Oct. 1980.

[56] J. Huang and P. Schultheiss, "Block quantization of correlated Gaussian random variables", *IEEE Trans. Commun. Syst.*, vol. CS-11, pp.289-296, 1963.