1987

Speech Signal Representation Through Digital Signal Processing Techniques

Debbie A. Kravchuk
University of Central Florida

SPEECH SIGNAL REPRESENTATION THROUGH DIGITAL SIGNAL PROCESSING TECHNIQUES

BY
DEBBIE ADA KRAVCHUK

B.S., FLORIDA ATLANTIC UNIVERSITY, 1981

RESEARCH PAPER

Submitted in partial fulfillment of the requirements for
the degree of Master of Science in Electrical Engineering
in the Graduate Studies Program of the
College of Engineering
University of Central Florida
Orlando, Florida

Summer Term
1987

ABSTRACT

This paper addresses several different signal processing methods for representing speech signals. Some of the techniques that are discussed include time domain coding which uses pulse-code, differential, or delta modulating methods. Also presented are frequency domain techniques such as sub-band coding and adaptive transform coding. In addition, this paper will discuss representing speech signals through source coding techniques via vocoders. A comparison of these different signal representation methods is included.

I would like to thank my husband Fred and my parents and family for all their patience, understanding, and encouragement. Their love and faith helped through the toughest times. I would also like to thank my friends for all their help and support.

I would like to acknowledge my employer, the Department of Defense, specifically, the Navy and Army. First, the Naval Training Systems Center, for selecting me for the Long Term Training program where I was sent to the University of Central Florida for one year. Second, the Department of the Army, Project Managers for Training Devices, for allowing me the time to complete this research paper.

Many thanks also to my committee members, Dr. Harden, Dr. Walker and my advisor Dr. Alsaka. Their patience, support, and guidence has made it possible for me to complete this research paper and fulfill all the requirements for the degree of Master of Science in Electrical Engineering.

PREFACE

Speech coding can be divided into two distinct areas:

1. Speech analysis.
2. Speech synthesis.

The study of these two areas requires an overall fundamental understanding of communication systems, the speech signal, and speech coding techniques.

This paper presents these concepts in the following manner: Chapter 1 is an introduction to the paper. Chapter 2 provides a general overview of a communication system. The speech signal is discussed in Chapter 3. Chapter 4 presents several common speech coding techniques. The last chapter, Chapter 5, provides examples of speech analysis and synthesis methods.

The reference material for this paper has come from sources that are concerned with signal processing. The majority of reference material originates from invited papers from IEEE Transactions in Acoustics Speech and Signal Processing. In addition, several signal processing textbooks were used to prepare this presentation. A complete list of the references used is provided at the end of this paper.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

A signal can be defined as a function that conveys information. The information is contained in a pattern of variation of some form which make up the signal [1]. For example, the signal could take the form of time or spatial varying patterns. The signal could also be represented mathematically as a function of one or more independent variables.

The independent variables are either continuous or discrete. Continuous signals are represented by continuous variable functions. These signals and functions are defined at all times within the variable range. Discrete signals are represented by discrete variable functions. Both are only defined at discrete times within the variable range. The continuous and discrete signals are also referred to as analog and digital signals, respectively.

Most often signals need to be processed before they are able to convey the informational content. Figure 1 [2] represents a general view of information transmission process.

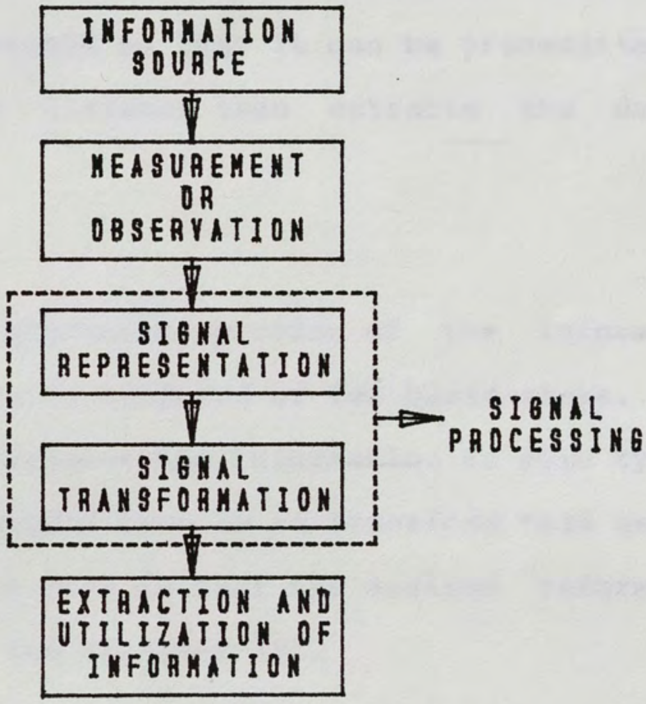


Figure 1. General view of information manipulation and processing.

The information or data lies in an information source or a type of data library. The data is measured or observed so that the desired information can be extracted. The information is then processed by first representing the information by a signal, then transforming the signal so that it could be conveniently transmitted. The last step is to extract the desired information from the processed signal representation. An example of this process is two people speaking on the telephone. The information source is a human speaker. The desired information is observed as speech as one person talks into the telephone. The speech is then signal processed so that it can be transmitted to the listener. The listener then extracts the desired information.

The signal processing portion of the information transmission process is composed of two basic steps. The first step is to represent the information in some type of general form. The second step is to transform this general representation into a form so that the desired information can be extracted by the receiver [2].

THE DIGITAL COMMUNICATION SYSTEMS

A communication system is composed of several main sub-systems. Figure 2 is a model of a typical communication system [3].

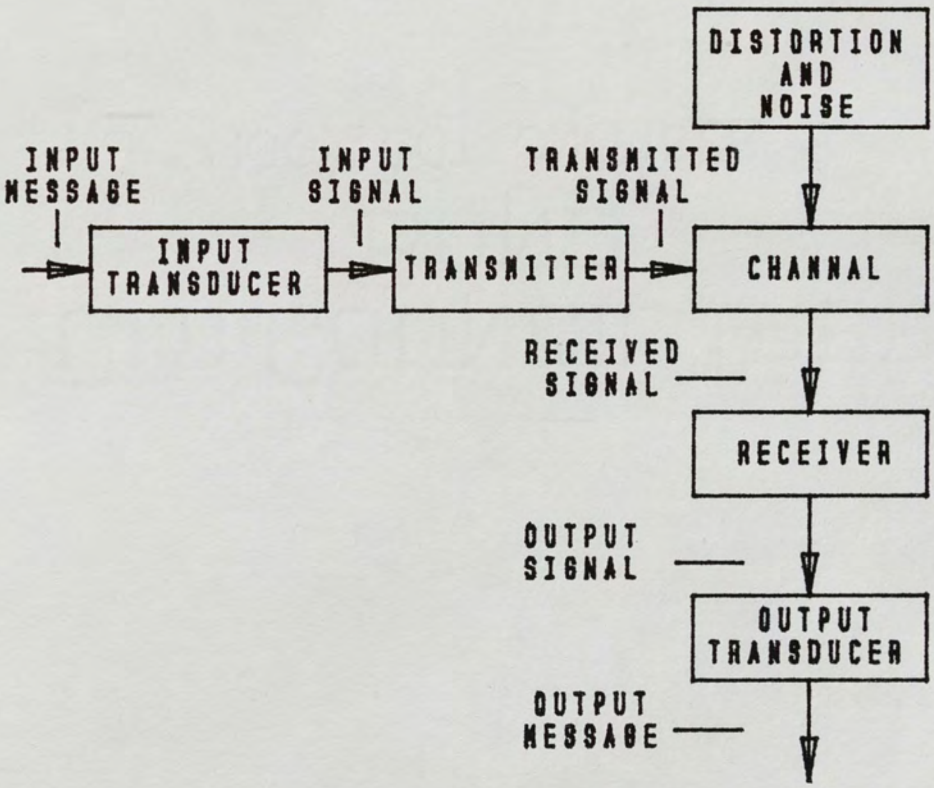


Figure 2. A communication system.

The communication system includes several basic components. An input message originates from a source. The information is then converted into an electrical waveform by the input transducer. The transmitter then modifies the message waveform so that it can be transmitted. The transmitter is composed of several subsystems such as a sampler, coder, quantizer and modulator. The output from the transmitter is then sent through a channel which could be a wire, coaxial cable, optical fiber, waveguide, etc. Distortion and noise are present throughout the entire communication process. Figure 2 illustrates the addition of distortion and noise while the signal is in the channel since this is where most of the contamination can occur. The modified message waveform is then received by the receiver. The receiver processes the modified message signal back into an electronic waveform by undoing the modifications that were done by the transmitter. The receiver is composed of several subsystems such as a demodulator, decoder, and filters. Finally, the receiver's output is fed to the output transducer where it is converted from an electrical waveform to the original signal or another desired signal.

Table 1 is provides a summary of the notation used in representing communication systems.

TABLE 1

COMMUNICATION SYSTEM NOTATION

SYMBOLS	EXPLANATION OF SYMBOLS
T	Highest frequency (in Hz) in the signal spectrum.
A/D	Analog - to - digital
L	Number of quantization levels.

Messages are of two types, digital or analog. Digital messages are composed of discrete values defined at discrete times. Where analog messages are composed of varying signal over a continuous range of time. A digital communication system can process a signal with greater accuracy than an analog system. Since a digital message is discrete values, it is easier for the receiver to extract the information from the distorted and noisy channel than it would be for an analog message. Currently, analog systems are widely being used, but with the cost reduction of fabricating digital circuitry, the analog systems are being phased out and replaced with the digital systems.

Many communication messages originate as analog signals and must first be converted to a digital signal so they can be transmitted through a digital system. This is done by analog-to-digital conversion or A/D conversion. A/D conversion is done by using the sampling theorem. The theorem states that a continuous signal can be reconstructed from samples of the signal if the samples are taken at a rate not less than 2T samples/second, where T (in Hz) is the highest frequency in the signal spectrum [3]. These sample values are still in a continuous range and must go through a process of quantization. This process approximates or rounds-off the sample to the nearest quantized or coded level. The signal is now digitized. The digitized signal is now an approximation of the analog signal. The approximation can become more accurate by increasing the number of quantization levels. The following are typical levels (L): for intelligibility of voice signals, L = 8 or 16, for commercial use, L = 32, and for telephone communication, L = 128 or 256 [3].

THE SPEECH SIGNAL

Digital signal processing is used to structure speech signals so that they can be analyzed or transmitted without losing the informational content. Before presenting models of speech signals, the natural speech production process should be discussed. Figure 3 [4] shows the basic parts of this process.

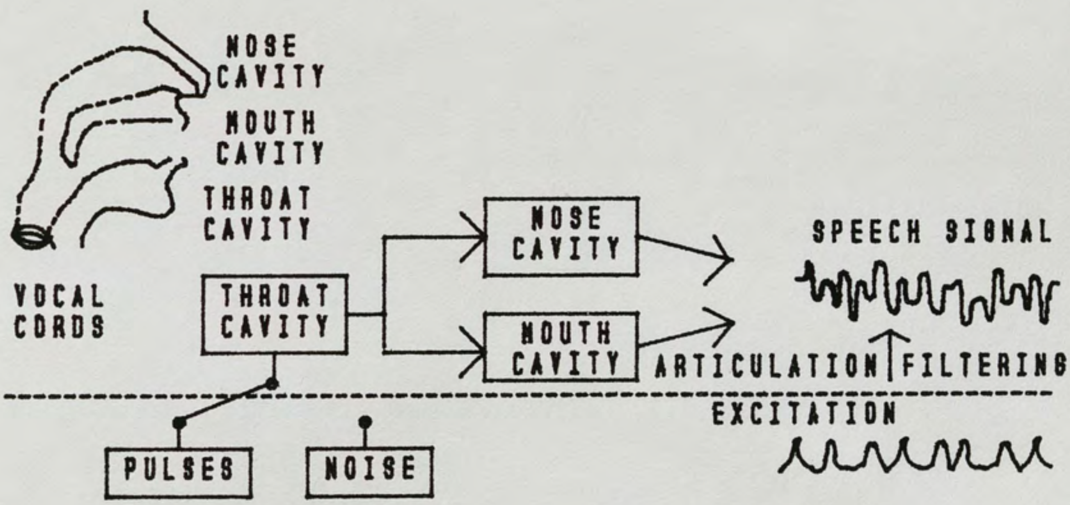


Figure 3. Principles of natural speech production.

A pulse excitation signal is produced by the vocal cords which are vibrating when the air stream from the trachea passes through. The pulses are then modulated within the cavities of the throat, mouth, and nose. The resulting signal will be a periodic voiced signal. The sound is defined by the filtering cavities. The properties of the cavities are defined by their geometric dimensions which can be changed during articulation. Figure 4 [4] is an overview of the speech signal.

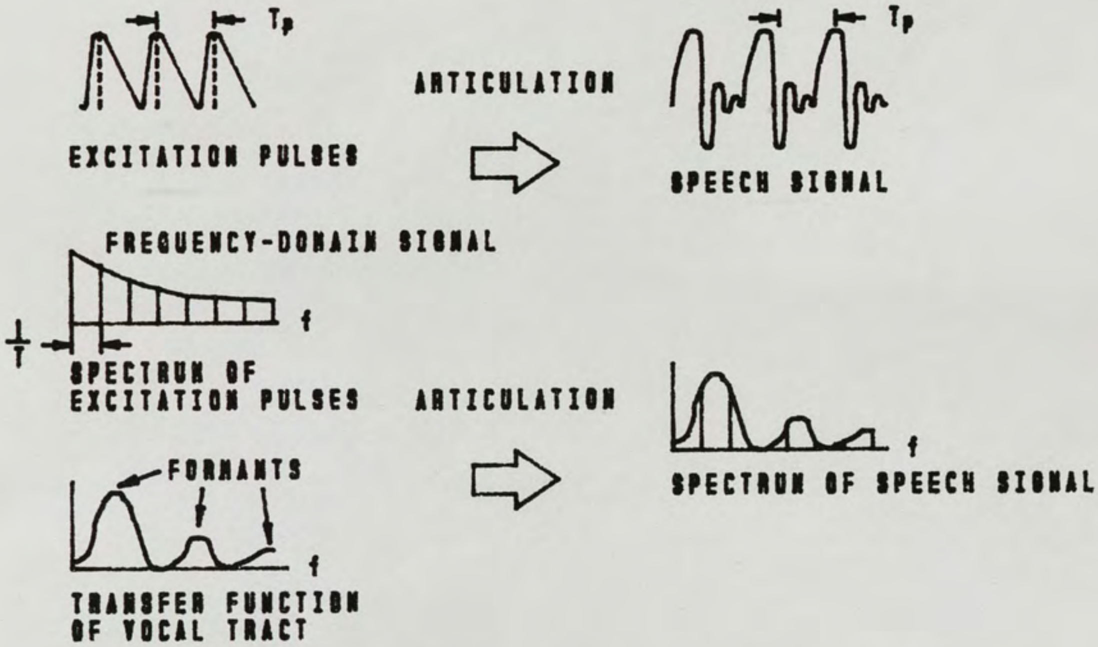


Figure 4. Time-domain and frequency-domain representation of speech.

For the time domain signal, the pitch of excitation pulse T is approximately 120 Hz in the average male voice. For the frequency domain signal, the excitation pulses of the signal have a spectrum that falls to high frequencies with about 6 to 10 dB/octave [4]. The spectral transfer function of the voice tract characterizes the sound of the speech signal.

A digital model of the vocal system is represented in Figure 5 [5]. Samples of the speech waveform, which correspond to the vocal sound, are the output to a time varying digital filter. The digital filter approximates the transfer characteristics of the vocal tract.

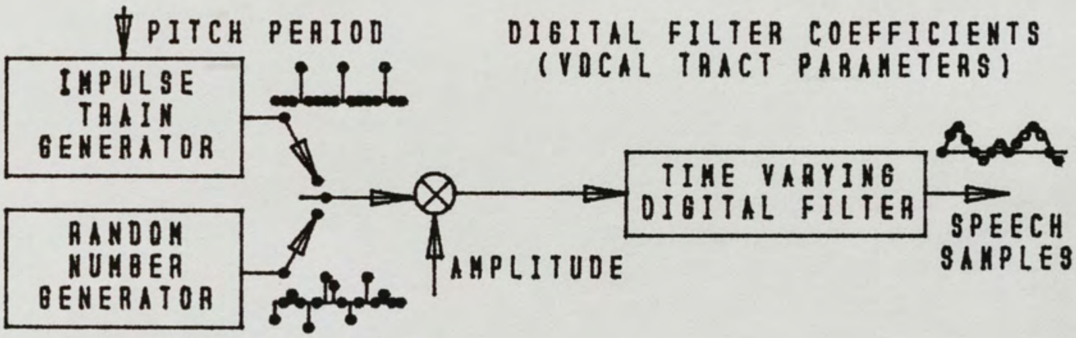


Figure 5. Digital processing model for the production of speech signals.

For a voiced signal, the digital filter is excited by an impulse train generator which corresponds to the excitation source in the vocal system. For an unvoiced signal, the filter is excited by a random noise generator which produces a flat spectrum noise [5].

SPEECH CODING TECHNIQUES

Speech coding is the process of digitally representing an analog speech signal so that it can be stored, transmitted, analyzed, synthesized or modified. Figure 6 [2] illustrates the general concept of coding speech signals.

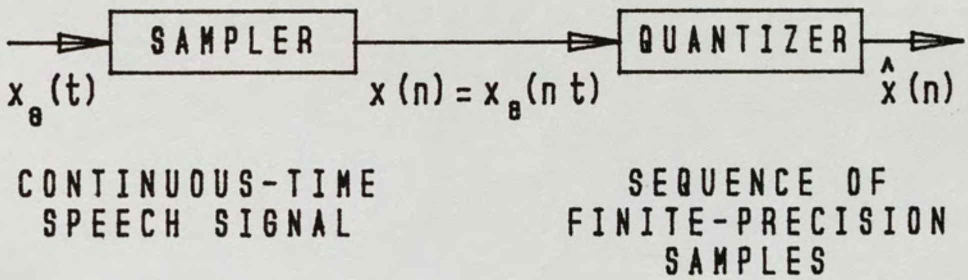


Figure 6. General block diagram depicting digital waveform representations.

Many types of coding methods are available. There are two major classifications of coders. First, is the waveform coder. This type of coder uses both time-domain or frequency-domain coding methods. The second is the source coder or vocoder. Most coding methods use either one of these methods or a combination of the two.

The waveform coder tries to get an exact digital reproduction or copy of the analog speech waveform. This type of coder is usually designed to be signal-independent. Therefore, the waveform coder can be designed to code a variety of signals including voiced, unvoiced, music, and tones [7].

A waveform coder design uses several fundamental speech properties and characteristics. Some of these properties include distributions for waveform amplitude and power, the non-flat characteristics of the speech spectrum (and equivalent autocorrelation functions), the quasi-periodicity of voiced speech, and the presence of silence or unvoiced sounds in a signal [7].

Some types of waveform coding methods for time domain coders include pulse-code modulation, differential pulse-code modulation, and delta modulation. For frequency domain coders, waveform coding methods include short time Fourier transforms, sub-band coding and adaptive transform coding.

The source coder method of signal coding depends on a prior knowledge about how the signal was generated at the source [7]. This method approximates some of the signal model and then extracts certain specific parameters so that the original signal can be characterized. Unlike waveform coding, source coding is not signal-independent. The input signal must be fitted into a specific mold and then parameterized [7]. This type of coding method is dependent on the signal; which is why it is referred to as source coding. Source coders for speech signals are usually referred to as "vocoders," which comes from "voice coders."

For the vocoder method, the source signal is assumed to be speech. This assumes that the signal is either voiced, generated from a quasi-periodic vocal-cord excitation, or unvoiced, generated from some random excitation. This type of source coder would extract such parameters such as voice pitch, the pole-frequencies of the modulating filters, and corresponding amplitude parameters [7]. Since the vocoder approximates the speech signal model, it can produce a digital representation of the speech signal at a much lower bit rate. However, the quality of the reconstructed speech signal is dependent on the speech signal model used.

Some types of vocoding methods include spectrum-channel, formant, pattern-matching and correlation vocoders.

Time Domain Coding Techniques

Digital techniques can be used to process, analyze, synthesize, and transmit speech signals. Representation of speech signals can be done with simple pulse code modulation techniques or with sophisticated vocoders. In this section, three digital coding techniques for processing the speech signal are presented. The coding is done by means of straightforward reconstruction of the speech waveform by discrete-time and discrete-amplitude representation. The techniques discussed are: (A) Pulse-Code Modulation, (B) Differential Pulse-Code Modulation, and (C) Delta Modulation. In addition, the adaptive versions of these methods are also considered.

The simplest digital representations of speech waveforms are concerned with direct representation of the speech waveform. Techniques as Pulse Code Modulation, Differential Modulation and Delta Modulation are all based on Shannon's Sampling Theorem which states that any bandlimited signal can be exactly reconstructed from samples taken periodically in time if the sampling rate is at least twice the highest frequency of the waveform. This is known as the Nyquist frequency. After a waveform has

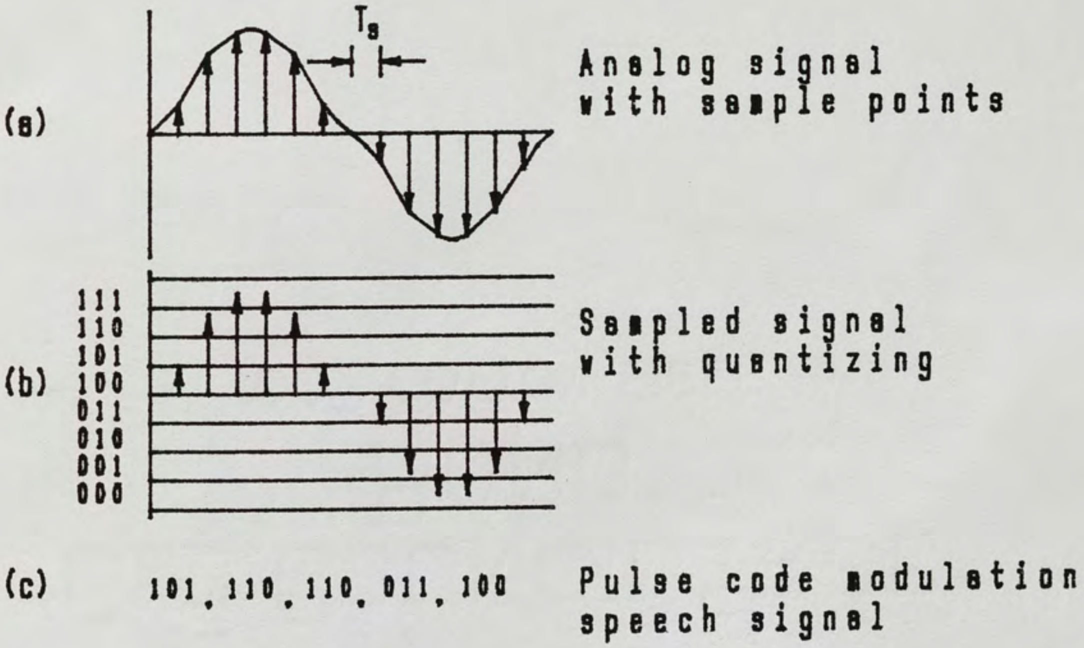


Figure 7. Pulse-code representation of analog signals.

been sampled at this frequency, the samples are then quantized or coded into a form of numbers. Figure 7 [4] represents an analog signal that has been sampled and quantized.

The following Table 2 [11] is provided as a summary of the notation used in presenting digital coding of speech waveforms.

TABLE 2

DIGITAL CODING OF SPEECH WAVEFORMS NOTATION

SYMBOLS	EXPLANATION
PCM	Pulse-code modulation
APCM	Adaptive PCM
DPCM	Differential PCM
ADPCM	Adaptive DPCM
DM	Delta modulation
LDM	Linear (non-adaptive) DM
ADM	Adaptive DM
Log- PCM	Logarithmic PCM
B	Number of Bits used to code a sample in PCM, DPCM (B= 1 for DM).
SNR	Signal-to-Noise (quantization error ratio).
Q	Constant step size in a non-adaptive quantizer.

TABLE 2 -- CONTINUED

p(a)	Probability density function of (a).
Y(r)	rth sample of quantizer output.
H(r)	Dimensionless quantity, proportional to Y(r), in a uniform quantizer.
M	Step size multiplier.
Q max, Q min	Maximum and minimum values of step size.
R	Dynamic range of adaptive quantizer.
Q opt	Optimum (constant) step size for a non-adaptive quantizer.
C	Correlation factor between successive samples.
Dr(j)	jth variance of the differences between samples.
X(r)	rth sample of quantizer input.
a(j)	jth predictor coefficient.
E(r)	rth sample of quantization error.
e(r)	rth sample of predictor error.
f	Sampling frequency.
b(r)	Binary code element at sampling instant r in DM.
K	Dimensionless constant in SNR expressions for DM.
F	Oversampling factor in DM. (F = 1 in PCM, DPCM)

Pulse Code Modulation

The method of Pulse Code Modulation (PCM) quantizes instantaneous sample values of the input speech waveform. Waveform coding by PCM involves the following steps [11]:

(1) The waveform is sampled at a rate of at least the Nyquist frequency.

(2) The amplitude of each signal sample is quantized into one of 2^B levels. This implies an information of B bits per sample.

(3) The discrete amplitude levels are represented by distinct binary words of length B. For example, with B=2, we can represent four distinct levels using code words 00, 01, 10, and 11.

(4) For decoding, the binary words are mapped back into amplitude levels, and the amplitude-time pulse sequence is low-pass filtered.

Assuming that steps (1), (3), and (4) can be implemented perfectly, we can determine the coder performance by the quantization error introduced in step (2). Figure 8 [11] illustrates an input signal, quantized output and the quantization error.

(a)

(b)

(c)

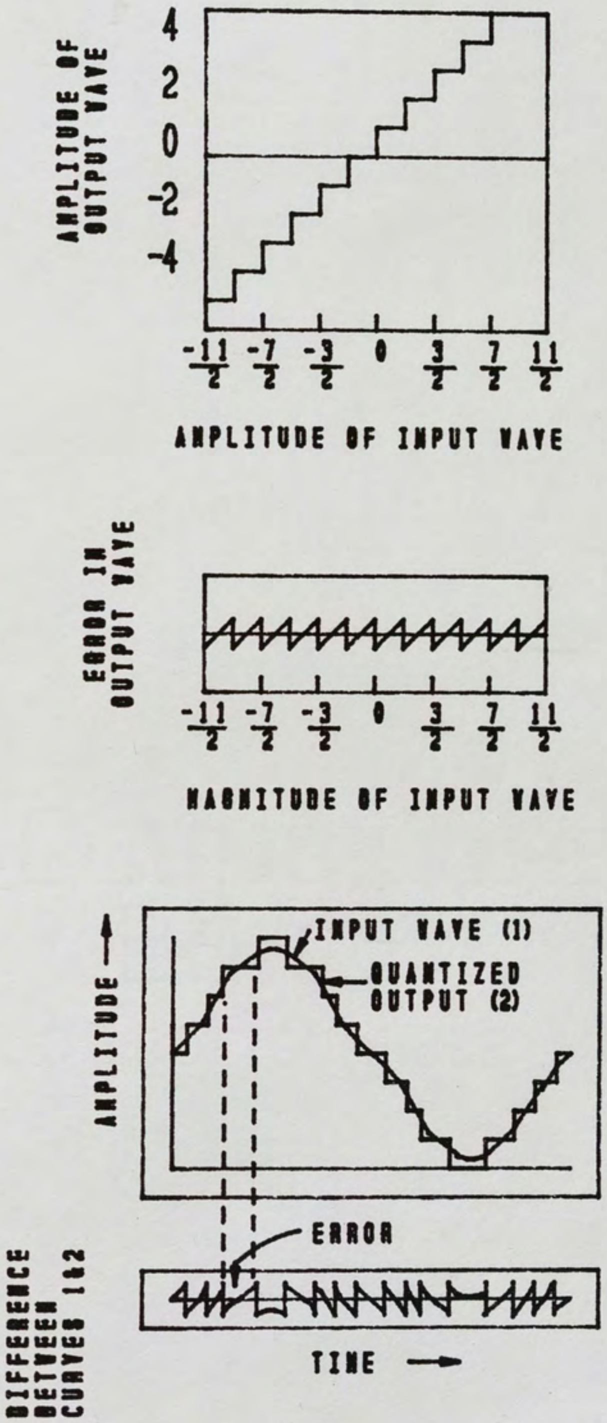


Figure 8. The quantizing principle.
(a) Quantizing characteristic.
(b) Characteristic of errors in quantizing.
(c) A quantized signal wave and the corresponding error wave.

To determine the quantizer's performance we must first define the signal to quantization error ratio (SNR).

Let Q = quantizer step size

then $p(E) = 1/Q \qquad (-Q/2) \leq E \leq (Q/2)$

=uniform distribution

and $$\int_{-Q/2}^{Q/2} E^2 \, p(E) \, dE = Q^2 / 12$$

= mean square value of quantization error

$$SNR = \frac{X^2_{rms}}{Q^2 / 12}$$

= signal-to-error ratio

Quantizers that modify their step size to match the signal variance are called Adaptive Quantizers (APCM). The idea is to work with the basic quantizer but modify its step size for each new input by a factor depending on the knowledge of which quantizer slots were occupied by previous samples [11].

Let the rth output of a B bit quantizer equal $Y(r)$. Then,

$Y(r) = [Q(r)/2][H(r)]$

Where H(r) is proportional to the output Y(r) of a uniform quantizer. Then the next step size is determined by

$$Q(r+1) = [Q(r)][M(|H(r)|)]$$

This step is determined by the last step size multiplied by a time invariant function of the magnitude of H(r) [11]. When the multiplier function is properly designed, the adaptive step size should match the step size at every sample to the updated estimate of the signal variance.

Figure 9 [11] illustrates the SNR characteristics of an adaptive and a non-adaptive quantizer. Where Q max and Q min represent practical constraints on the adaptation logic. The ratio of (Q max)/(Q min) determines R, the dynamic range of the input variance, which the quantizer can handle. The quantity Q opt is the optimum (constant) step size for a non-adaptive quantizer and Q mid is the range in which a given input signal can tolerate for a specified minimum performance (SNR). As you can see, an adaptive quantizer can increase the dynamic range.

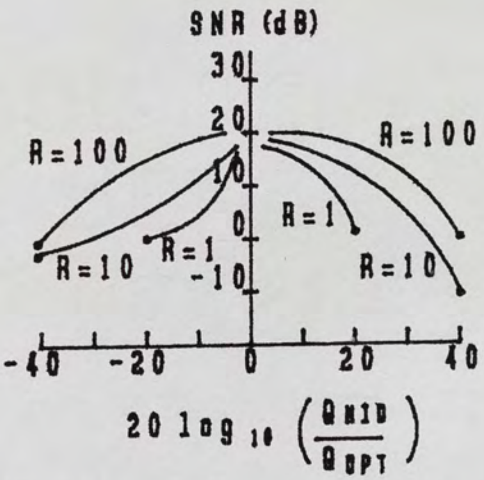


Figure 9. Companding characteristic of adaptive quantizer.

In addition to the APCM, another solution to the mismatch problem is the use of a non-uniform PCM. A non-uniform quantizer has a changing step size, whereas a uniform quantizer's step size is constant. An example of a coder with an increasing step size is a logarithmic quantizer or Log-PCM. This type of coder has the advantage of taking large end steps without increasing the total number of quantization levels. However, the disadvantage of a lower SNR may occur using this method [11].

A coding system employing a uniform quantizer can be designed to perform as a Log-PCM system. This can be done by compressing the input signal to the uniform quantizer and then expanding the coded output.

Differential Pulse Code Modulation

The basic principle of Differential Pulse Code Modulation (DPCM) is to quantize the changes in the signal rather than instantaneous sample values as in Pulse Code Modulation (PCM). Speech that is sampled at the Nyquist rate has a correlation factor (C) between successive samples. To define the SNR, let

$$C = [\ \langle [X(r)][X\ (r-1)] \rangle\ ]\ /\ [\ \langle X^2(r) \rangle\ ]$$

and the variance of the first difference is given by

$$\begin{aligned} \langle Dr^2 (1) \rangle &= \langle\ (X(r)-X(r-1))^2\ \rangle \\ &= \langle\ X(r)^2\ \rangle\ [2\ (1-C)] \end{aligned}$$

for C > 0.5, Dr(l) has a smaller variance than the signal variance. Therefore, it may be advantageous to code values of the signal difference, rather than the signal itself, then use an integrator to reconstruct the signal from the quantized values of Dr(l). The quantization error is proportional to the variance of the signal present at the input of the quantizer. So if we input Dr(l) instead of the signal, we can get a better SNR.

The DPCM can be used to quantize the difference between input X(r) and its linear prediction. In Figure 10 [11], feedback is used so that quantization errors are not accumulated. Another important note is that this type of coding is independent of the number of bits used.

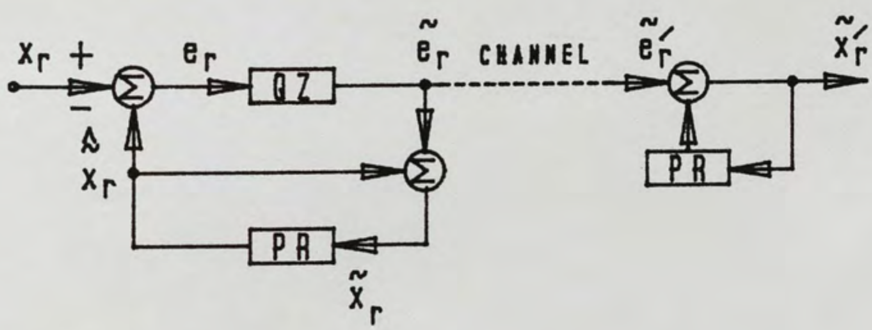


Figure 10. Differential PCM. QZ: quantizer; PR: predictor:

$$\hat{\tilde{X}}(r) = \qquad a(j)\ \tilde{X}(r-j).$$

If the channel is error-free,

$$\tilde{X}(r)' = \tilde{X}(r)$$
$$= \text{reconstructed signal.}$$

There are two different types of adaptive DPCMs. The first is the adaptive predictor. This ADPCM adapts predictor coefficients to the changing spectral properties of speech signals. It does this by storing finite sections of speech, calculates the autocorrelation function for the section, then determines the optimum predictor vector.

The second type of ADPCM is the adaptive quantizer. The adaptive quantizer used in the ADPCM is basically the same as in the APCM. The adaptive quantizer uses quantizer memory of previous samples to match the quantizer step size to the input. The input is the variance of the predictor error to the DPCM.

Delta Modulation

Delta Modulation is a one bit version of the DPCM in which oversampling of the signal is used to increase the correlation C between samples. A Delta Modulator approximates an input time function by a series of linear segments of constant slope. This type of coder is usually called a Linear Delta Modulator, LDM.

The LDM uses staircase approximations to a bandlimited input signal sampled at a rate of f (where f is much higher than the Nyquist frequency). Figure 11 [11] illustrates a linear Delta Modulator.

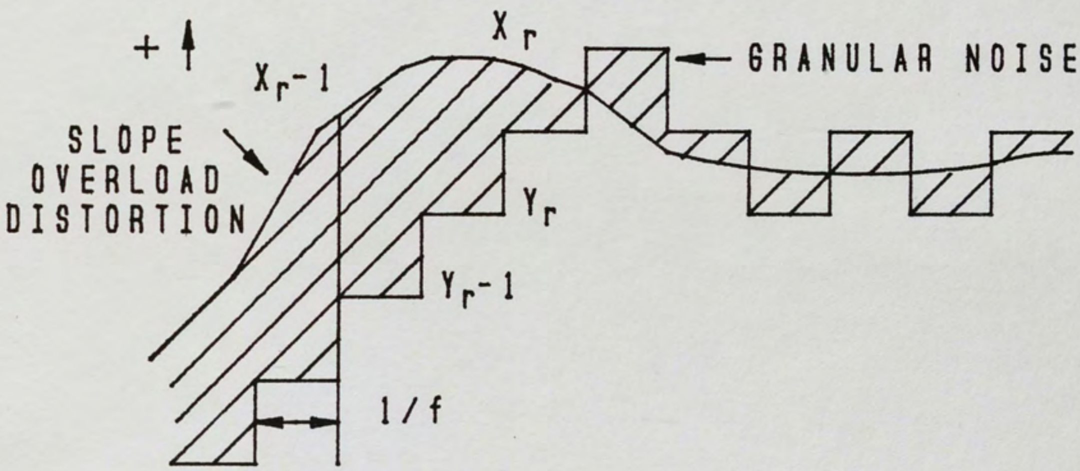


Figure 11. Linear delta modulation (LDM)

The basic DM principle can be shown in the following two equations:

$$b(r) = \text{sign}\ [\ X(r) - Y(r-1)\ ]$$

$$Y(r) = Y(r-1) + [\ Q\ ]\ [b(r)]$$

In the Delta Modulator, b(r) represents a binary code element. It is the sign of the difference between input X(r) and the latest staircase approximation Y(r-1). Y(r) is then incremented by step size Q in the direction of b(r).

Figure 11 [11] also shows two different types of quantization error associated with the LDM. The first is slope overload. This error occurs when the step size is too small to follow a steep segment of input waveform.

The second type of error is granular noise. This type of error occurs when the step size is too large for a relatively flat input waveform.

The block diagram of a LDM is presented in Figure 12 [11]. This diagram is very similar to the DPCM. The important differences are the use of a one bit quantizer in the DM and the replacement of the general predictor with an integrator [11].

The SNR is defined by:

$$SNR = [K][F]\ /\ [1-C(F)] \qquad F>>1$$

where K is a dimensionless quantity of the order of unity.

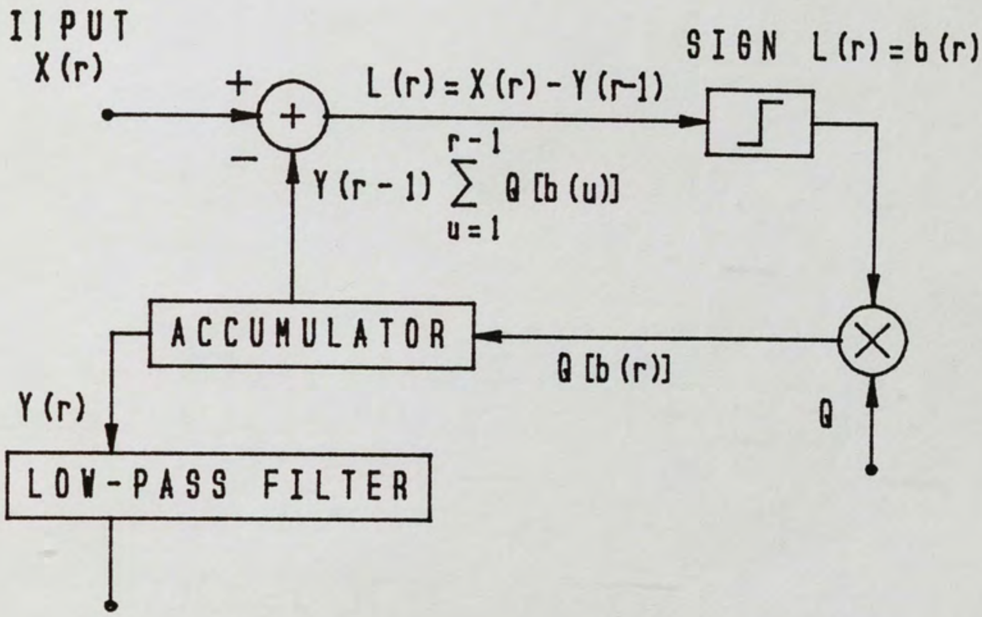


Figure 12. Block diagram of LDM

F = oversampling index

$$= \frac{\text{sampling rate}}{(2)\ (\text{highest frequency of the waveform})}$$

C = correlation between successive samples of the input

The adaptive Delta Modulator is the same as for the APCM and ADPCM. The ADM uses adaptive techniques to improve dynamic range by varying step size.

The step size is modified to increase over steep segments and decrease over relatively flat segments of input waveform.

Comparison Of Direct Waveform Coding Techniques

The quantization performance for time domain coders can be expressed by the signal-to-quantization error ratio or signal-to-noise SNR ratio. Figure 13 [11] illustrates the SNR as a function of bit rate for the following three coders: 1.) logarithmic PCM, 2.) ADPCM with adaptive quantizer and a simple first order predictor, 3.) ADM with a one-bit memory. The results used a computer simulated male speech signal of "A lathe is a big tool" [11]. Figure 13a considered a bandwidth of 200-3200 Hz and Figure 13b 200-2400 Hz. The sampling frequencies (f) for PCM and ADPCM were 8 kHz and 6.6 kHz, respectively. The number of bits per sample (B) was variable, and the bit rate was determined by the product of Bf. For the ADM the bit rate equaled the sampling frequency which was variable. The following comparisons can be made from the plots [11]:

a) The ADPCM has a constant gain over the PCM, 12 dB in Figure 13a and 8 dB in Figure 13b. The gain is due to differential encoding technique of the ADPCM which is independent of B.

b) The ADM and PCM are both dependent on the bit rate. The ADM SNR increases as the cube of the bit rate, and the PCM increases exponentially. The ADM has a better SNR at lower rates, but as the rate increases it saturates. As seen in figures 13a and 13b, the cross over points in bit rates are 50 kbits/sec and 30 kbits/sec, respectively.

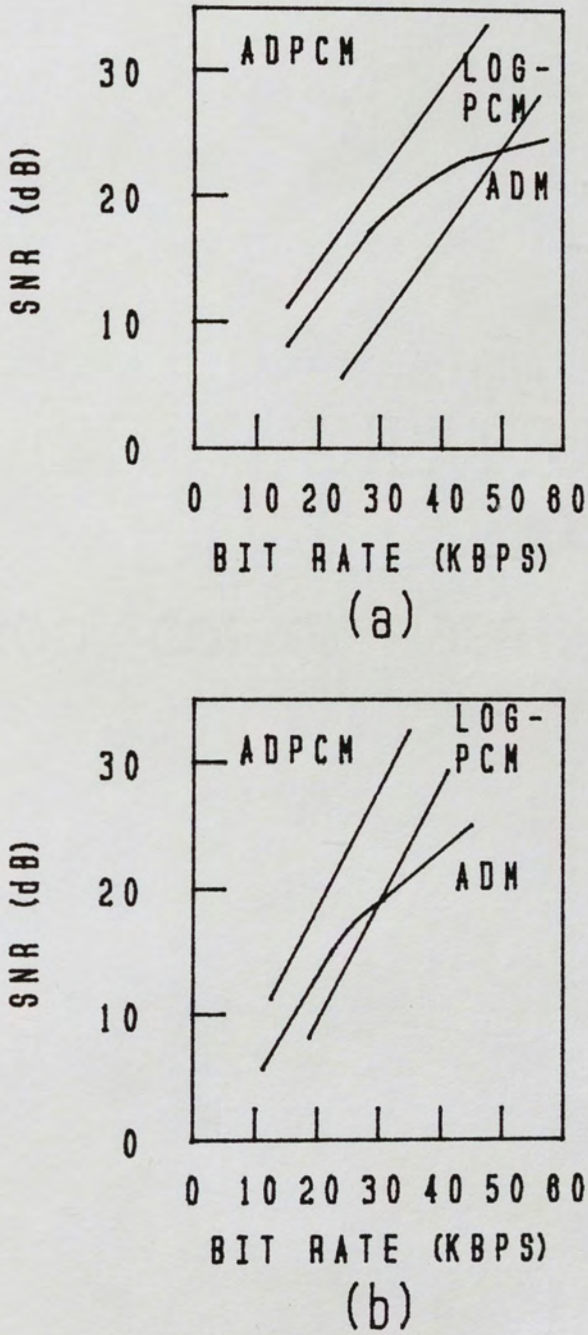


Figure 13. Comparison of SNR versus bit rate functions.
(a) Bandwidth: 200-3200 Hz.
(b) Bandwidth: 200-2400 Hz.

Below these cross over points the ADM performs better than the logarithmic PCM.

c) The ADPCM appears to have the best SNR performance. The SNR performance is only slightly higher at lower bit rates. This slight increase in SNR should be weighed against the added complexity of a multi-bit quantizer.

This comparison only defines coder performance as a function of SNR. Other performance parameter which may be considered include environmental tolerances, vibration tolerances, reliability. In addition, other factors such as cost, availability, military or commercial standards, size and weight, must be considered when selecting a coder. In general, a coder overall performance will always depend on its application.

Frequency Domain Coding Techniques

In the past, speech coders have usually fallen into one of two classes. One is the waveform coder which tries to reproduce the original speech waveform from samples of the input speech waveforms. The second is vocoders which extract certain parameters of the input waveform and then fit them into a speech model for reproduction. Recently, a third class of speech coders has been developed called frequency domain coders. The frequency coder divides the speech signal into a set of frequency components which are then separately encoded using some techniques of time domain waveform coding. This technique allows each band to be encoded by the desired method required for that band.

Two basic types of frequency domain coders will be presented. The first is sub-band coders. This type of coder uses filter bank analysis to separate the speech spectrum into typical 4-8 contiguous sub-bands [13]. The second type of frequency coders is transform coders. In this case, the coder does a block by block transform analysis to separate the signal into typically 512 frequency components [13]. Short-time spectral analysis of the input signal is used in both coders. A brief overview of this analysis is provided below.

Short-Time Fourier Transform

The main characteristic of a frequency domain coder is that it extracts certain signal parameters or components from a short-time spectrum for encoding. To do this the coder must use short-time Fourier transform techniques. The following Table 3 is provided as a summary of the notation used in discussing short-time Fourier transforms.

The short-time Fourier transform of the sequence x(n) is defined by:

$$X_n(e^{jw}) \triangleq \sum_{m=-\infty}^{\infty} h(n-m)x(m)e^{-jwm}$$

TABLE 3

SHORT-TIME FOURIER TRANSFORM NOTATION

SYMBOLS	EXPLANATION
$x(m)$	mth sample of input signal.
$h(n-m)$	window which reflects the portion of $x(m)$ to be analyzed, where n is the discrete time index.
w	continuous frequency.

The short-time Fourier transform, which is a time-dependent transform, is a function of two variables, time (n) and frequency (w). This can be used to represent either the filter bank analysis or block transform methods.

For the filter bank analysis representation, the frequency is fixed at w=W. The filters output is seen as the output of a linear time-invariant filter having an impulse response h(n) which is excited by a modulated signal [13]. This is expressed as,

$$X_n(e^{jW}) = h(n) * [x(n)e^{-jWn}]$$

where (*) is the convolution operation. In this case, h(n), referred to as the analysis filter, determines the bandwidth of the analysis around the center frequency W of the signal x(n) [13].

For the block Fourier transform representation, the time index n is fixed at n=N. The resulting analysis is the normal Fourier transform as seen below,

$$X_N(e^{jw}) = F\{h(N-m)x(m)\}$$

where F{ } is the Fourier transform operation. The time width of the analysis around the fixed time N is determined by the analysis window, h(N-m) [13].

Inverse short-time Fourier transform techniques are used to recover the original signal x(n) from the short-time spectrum analysis. The synthesis equation is given as,

$$\hat{x}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \sum_{r=-\infty}^{\infty} f(n-r)X_r(e^{jw})e^{jwn}\,dw$$

where f(n) is referred to as the synthesis filter or synthesis window [13]. For the synthesized signal to equal the original signal, the following relationship must also be true [1]:

$$\sum_{n=-\infty}^{\infty} f(-n)h(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} F(e^{jw})H(e^{jw})dw$$

As in the analysis, the short-time Fourier synthesis equation can be used for the filter bank and block transform representations.

For the filter bank representation f(n) has the following form,

$$f(n) = \delta(n) \;/\; h(n), \qquad h(n) \neq 0$$

and the synthesis equation has the following form.

$$\hat{x}(n) = \frac{1}{2\pi h(0)}\int_{-\pi}^{\pi} X\,(e^{jw})e^{jw}\,dw$$

In the case of the block transform representation, f(n) has the form

$$f(n) = 1 \;/\; H(e^{j\theta}) \qquad \text{for all n.}$$

Sub-Band Coding

Sub-band coding systems, which are based on wide-band analysis and synthesis, can best be described by the filter bank interpretation method. This interpretation is illustrated in Figure 14 [13] for a M channel filter bank. The signal analysis is done by modulating the center frequency of each frequency band to dc. Each band is then low-passed filtered with h(n) and compressed in time by a factor R:1. To synthesize the signal, the sub-band signals are expanded in time, by filling in with zeros, and increased by a factor of 1:R. The signals are then filtered with f(n) and modulated so that each band's center frequency is back to its original location [13]. All these sub-bands are then summed together to produce the synthesized output.

The filter band interpretation is only a concept used to understand sub-band coding. In reality, the sub-bands are usually implemented as a low-pass translation of a frequency band to dc in a method similar to single-side-band modulation so that the signals are real instead of complex as seen in Figure 14. Figure 15 illustrates the general procedure for this process.

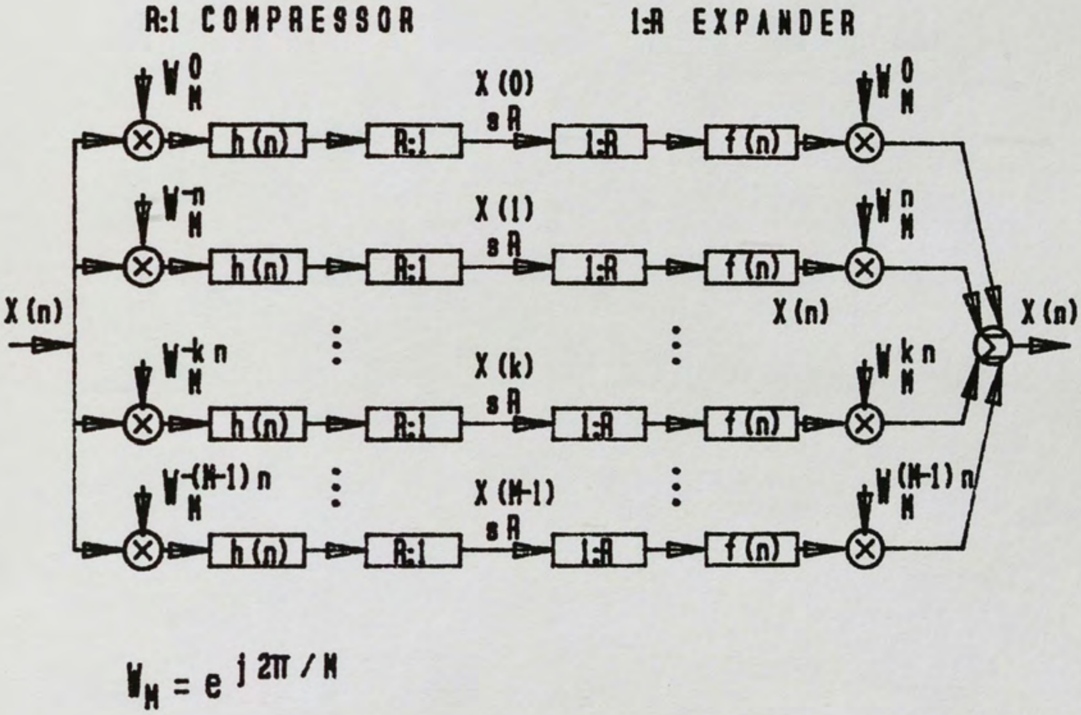


Figure 14. Filter bank interpretation of short-time analysis/synthesis.

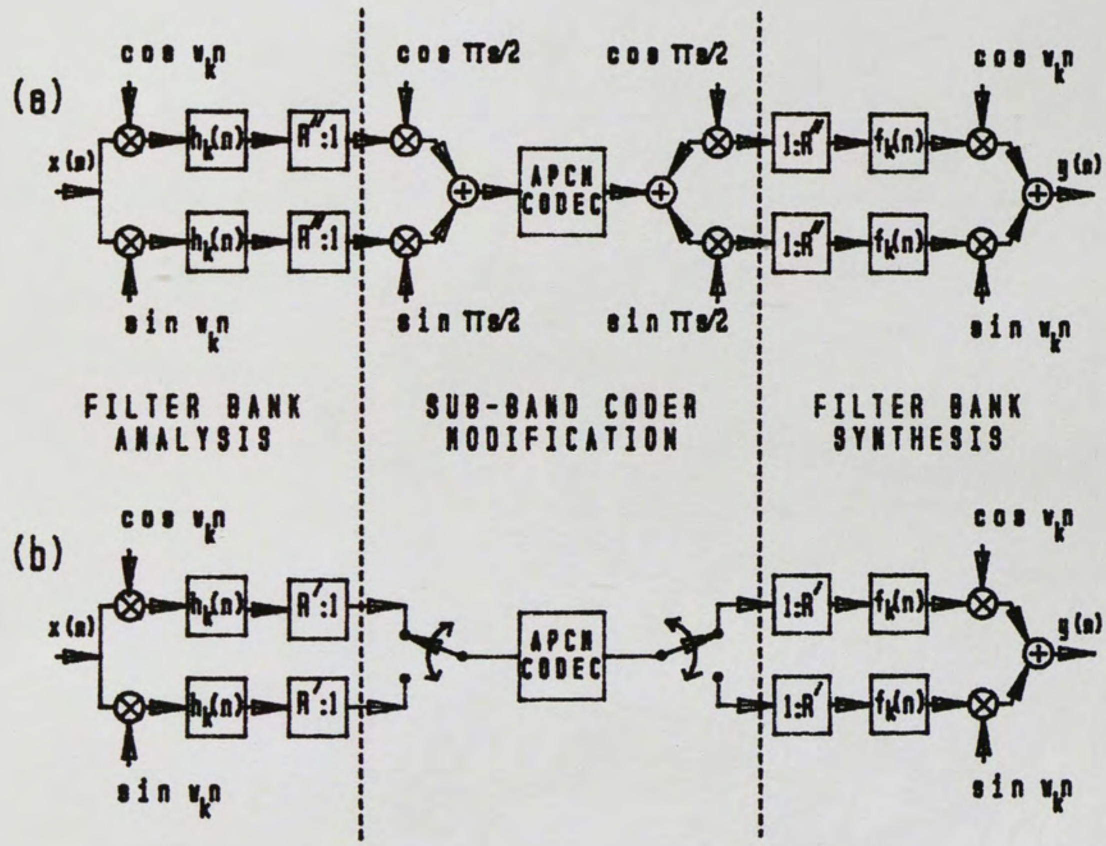


Figure 15. Sub-band Coder:
(a) block diagram of signal processing operations for sub-band signals.
(b) A simplified interpretation showing $y_k(n)$ and filter bank outputs.

Adaptive Transform Coding

Transform coding systems, are based on narrow-band analysis and synthesis. This type of system can be represented by the block transform interpretation method as seen in Figure 16 [13]. The input signal is divided into m different segments with each segment then being windowed by the analysis window. An M point Fourier transform is then applied to transform each windowed time segment into a sampled short-time spectrum. To reproduce the signal, the inverse discrete Fourier transform is applied to each sampled short-time spectrum to obtain its short-time time domain representation. The synthesis window f(n) filters across the overlapping short-time segments to reproduce the time signal [13].

The block diagram of an adaptive transform coder is illustrated in Figure 17 [13]. The input speech signal is buffered into short-time blocks and then transformed. The block components are then adaptively quantized and sent to the receiver. The receiver then decodes and inverse transforms these blocks. These new blocks are then used to produce the output signal. The side information is used by the step size adaptation and bit allocation [13].

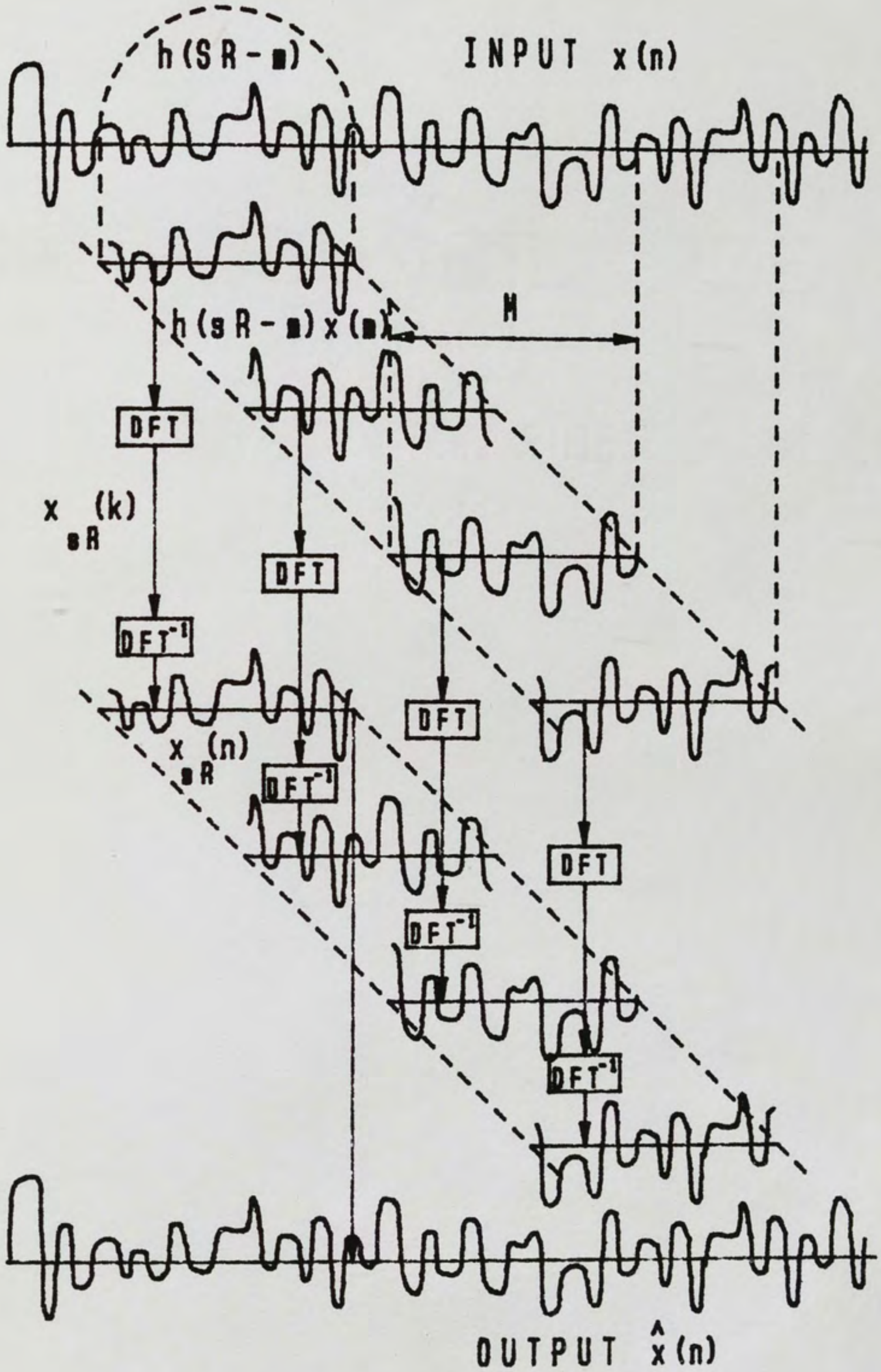


Figure 16. Block transform interpretation of short-time analysis/synthesis.

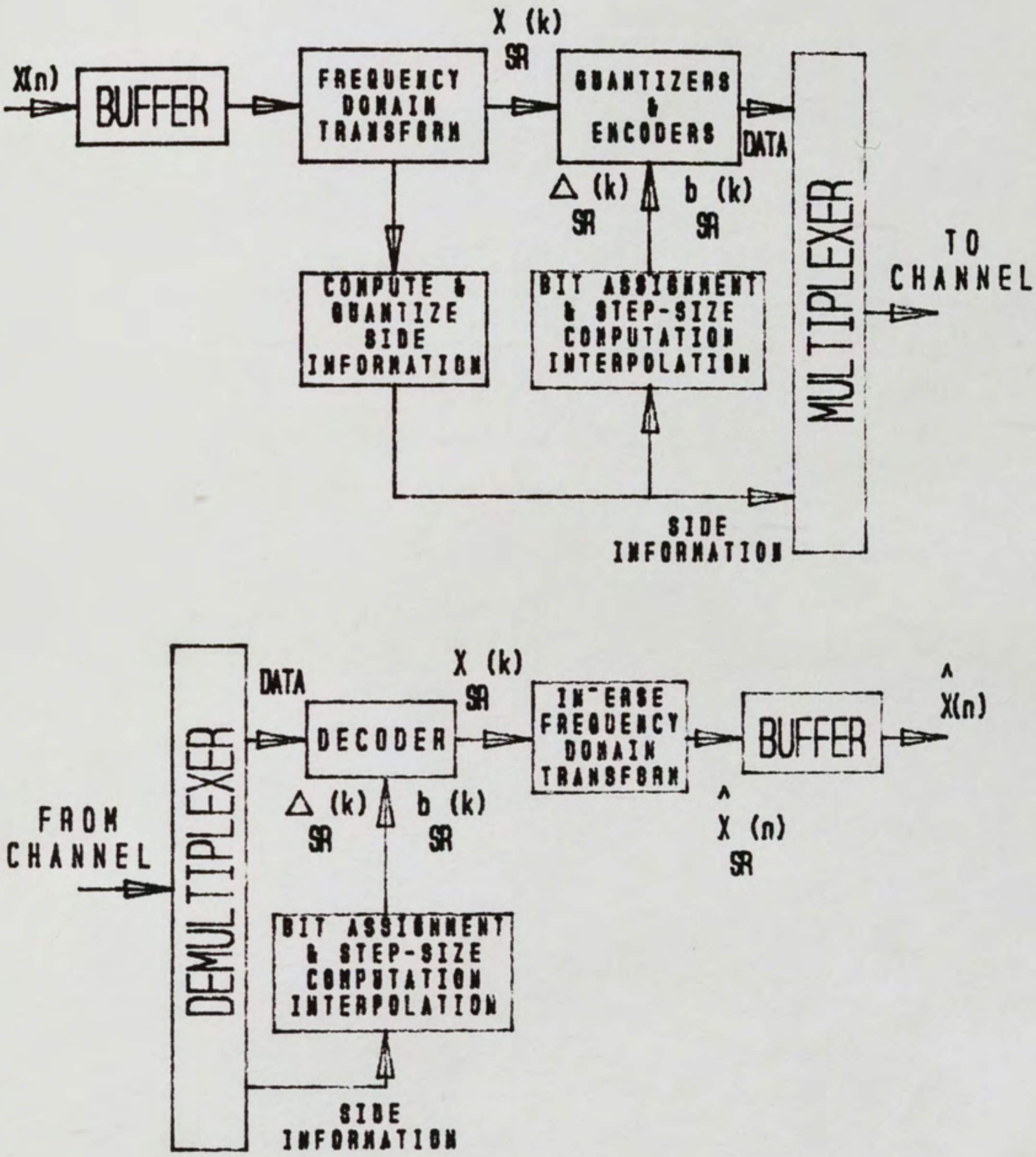


Figure 17. Block diagram of adaptive transform coder.

The following four major signal processing operations must be considered in adaptive transform coding [13]: 1) analysis and synthesis operations and the choice of the transform, 2) the spectral parameterization operations, 3) the step size adaptation and bit allocation, 4) and the quantization and multiplexing of the signals.

Vocoders

Vocoders are used to analyze and synthesize speech. The term vocoder was derived from the words voice and coder as defined by Homer Dudley in 1939. The vocoder is a speech coding device with many applications in the transmission, storage, and encryption of speech signals.

Spectrum-Channel Vocoder

The spectrum-channel vocoder is one of the oldest methods used for speech analysis-synthesis. In this type of vocoder, the spectral envelope, Figure 18 [18], is typically represented by 10 to 20 samples spaced along the frequency axis. The spectral fine structure, also shown in Figure 18, is represented by an additional parameter which measures the fundamental frequency of voiced sounds. For unvoiced sounds and silence, it is equal to zero.

A block diagram of the spectral-channel vocoder is shown in Figure 19 [18]. The speech signal is separated into fourteen different spectral bands covering the frequencies from 200 Hz to 3200 Hz. This is a typical frequency range for telephone signals [18]. Each of the bands has a bandwidth between 100 Hz and 400 Hz. The output of each filter is passed through a rectifier then low-pass filtered. These outputs represent a time-varying average signal amplitude for each spectral band. These fourten outputs represent the short-time spectrum of the speech signal.

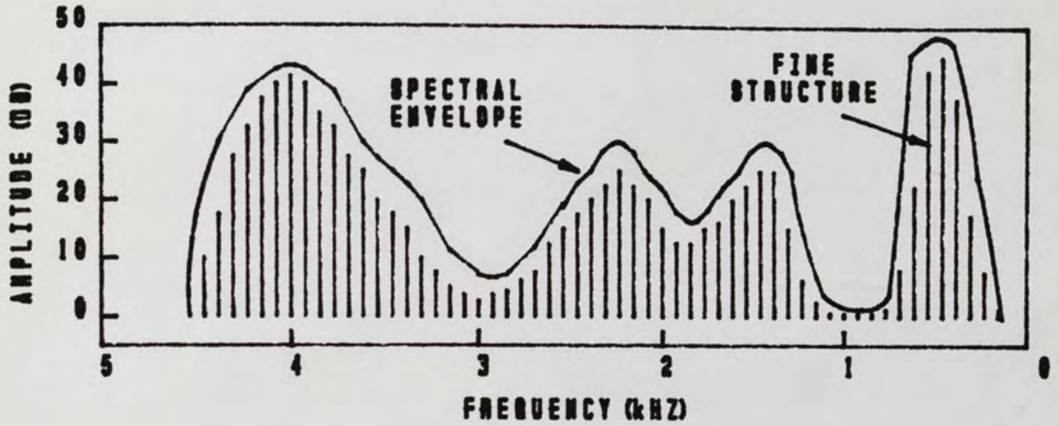


Figure 18. Spectral envelope showing four peaks or formants. The fine structure corresponds to a fundamental frequency of about 110 Hz.

Included in the block diagram is a voiced-unvoiced detector and a pitch detector which determines the fine structure of the speech signal and produces a corresponding narrow-band signal.

The vocoder synthesizer recovers the original signal. The narrow-band signals, transmitted from the analyzer, are input into time-varying filters (consisting of modulators and narrow band-pass filters). The input to the time-varying filter is a flat spectral excitation signal which has the spectral fine structure of periodic pulses for voiced speech sounds or "white" noise for unvoiced sounds [18].

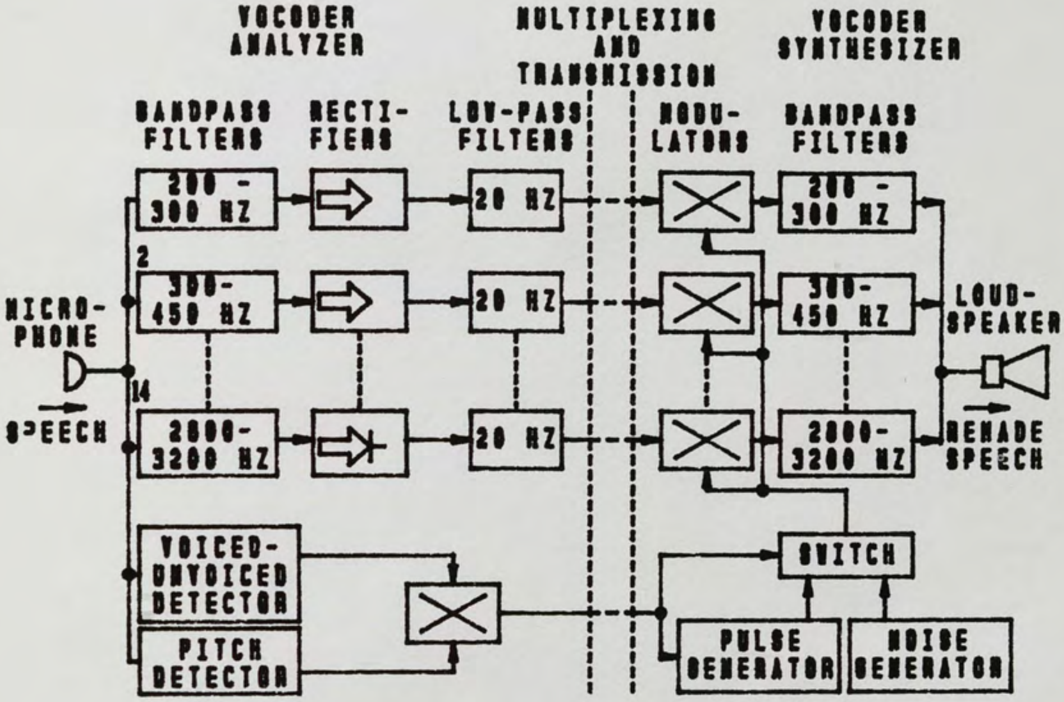


Figure 19. Block diagram of a spectral channel vocoder.

Formant Vocoder

Another type of vocoder is the formant vocoder. The term formants, which was derived from musicology, represents frequency regions that are characterized by several prominent maxima [18]. These regions represent the resonances of the vocal tract. For adult speech, three formants are typically used below 3000 Hz.

The analyzer in a formant vocoder attempts to determine the frequency locations of the major formants. These locations are then transmitted and used to control the resonance of the formant synthesizer which uses three or more single tuned resonance circuits [18].

A formant vocoder can also use a parallel synthesizer as shown in Figure 20 [18]. This type of vocoder extracts both the formant amplitudes and the frequencies from the speech signal. This information is then used by the synthesizer to control the modulators and the tunable resonators [18].

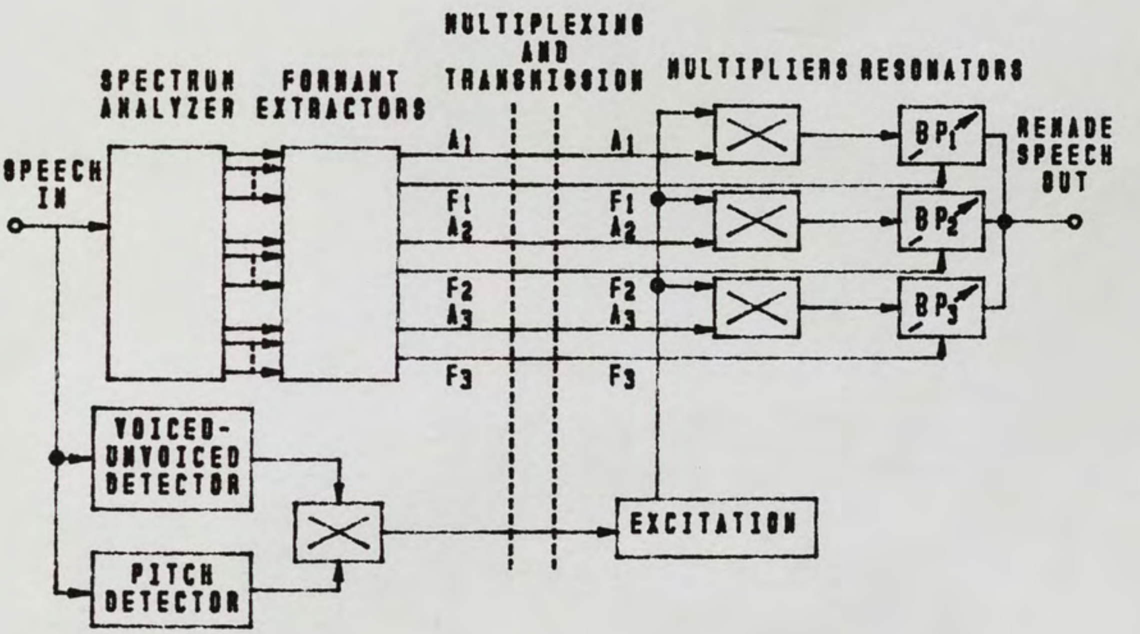


Figure 20. Formant vocoder using parallel synthesis. Formant amplitudes and frequencies are extracted in the analyzer and used to control the modulators and the variable parallel filters in the speech signal to produce an artificial speech signal.

The formant vocoder can provide very realistic and natural speech when the synthesizer is supplied with the proper control signals. The difficulty comes when trying to automatically extract these control signals with a high degree of accuracy.

Pattern-Matching Vocoder

The pattern-matching vocoder is similar to the formant vocoder in that it also uses spectral information. The pattern-matching vocoder determines the best match between the short-time speech spectrum and a set of previously stored spectral patterns. This best match is then transmitted to the synthesizer where it is used to reproduce the original speech signal [18].

Correlation Vocoder

The autocorrelation function can be used to represent the power spectrum of a signal. A similar relationship can also be used to represent a speech signal by a time-varying, short-time autocorrelation function instead of its short-time spectrum [18]. The autocorrelation vocoder uses this type of speech representation.

The speech signal is entered into the analyzer of the autocorrelation vocoder. The signal is equalized then multiplied by various delayed samples of itself as shown in Figure 21 [18]. The products are then low-pass filtered to form correlation channel signals [18].

These time-varying signals represent a short-time autocorrelation function. This representation is then used to synthesize the speech signal back into the time-domain. The time-domain synthesizer is also shown in Figure 21. The synthesizer impulse responses are controlled by the autocorrelation signals. The excitation signal applied to the synthesizer has a flat spectral envelope and the proper fine structure (quasi-periodic pulses for voiced speech sounds and "white" noise for unvoiced sounds) [18].

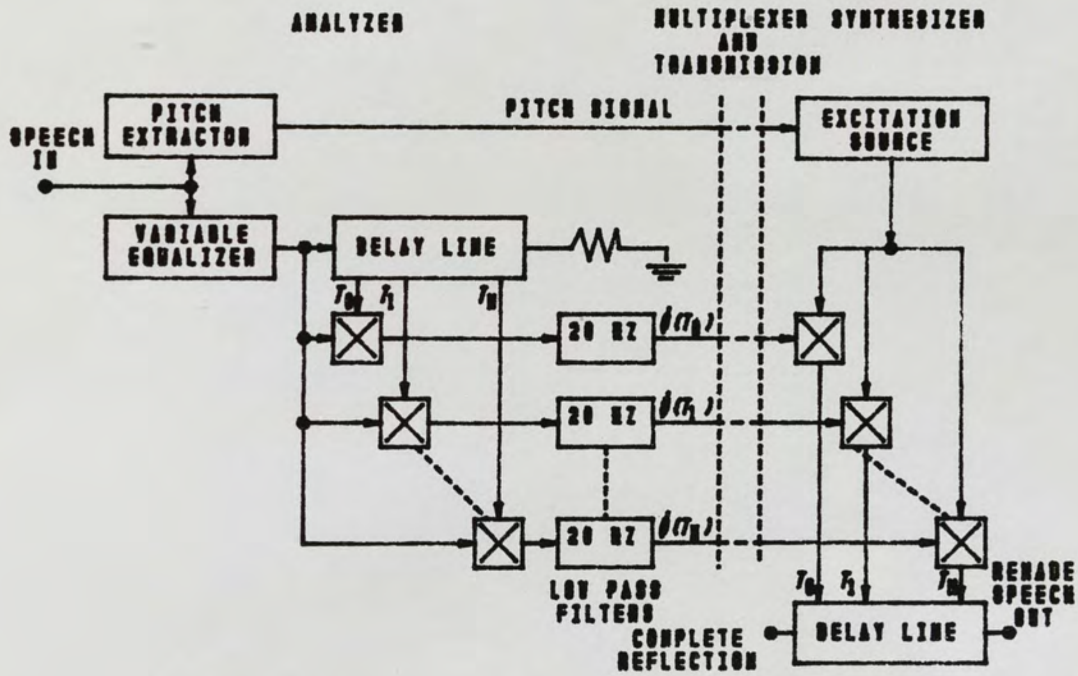


Figure 21. Block diagram of an autocorrelation vocoder.

In the autocorrelation vocoder, the spectral envelope of the synthesized output is equal to the spectral envelope of the autocorrelation function, which in turn can be related to the power spectrum of the original speech signal. A problem can occur with the autocorrelation vocoder since the output signal spectrum is the square of the input signal spectrum. All level differences in the original speech signal are doubled which causes weaker formants to be suppressed when compared to stronger formants [18].

The spectrum squaring problem can be eliminated by using several different methods. One method involves using special equalizers that perform spectral squaring operations. The equalizer divides the speech signal into several frequency bands. The average amplitude is determined for each band and then square-rooted. These signals are then recombined into one equalized signal [18]. After the spectrum squaring process, the output will have the approximate spectrum as the original speech signal.

Another method used to avoid the spectral squaring problem in the autocorrelation vocoder is cross-correlation analysis. The cross-correlation analyzer is similar to the autocorrelation analyzer except that it uses a spectrum flattener. The speech signal is passed through the spectral flattener so that the spectral envelope has been equalized to be nearly constant [18]. The signals are then synthesized as before producing signals that approximate that original spectral envelopes instead of its square. These two methods are shown in figures 22 [18] and 23 [18].

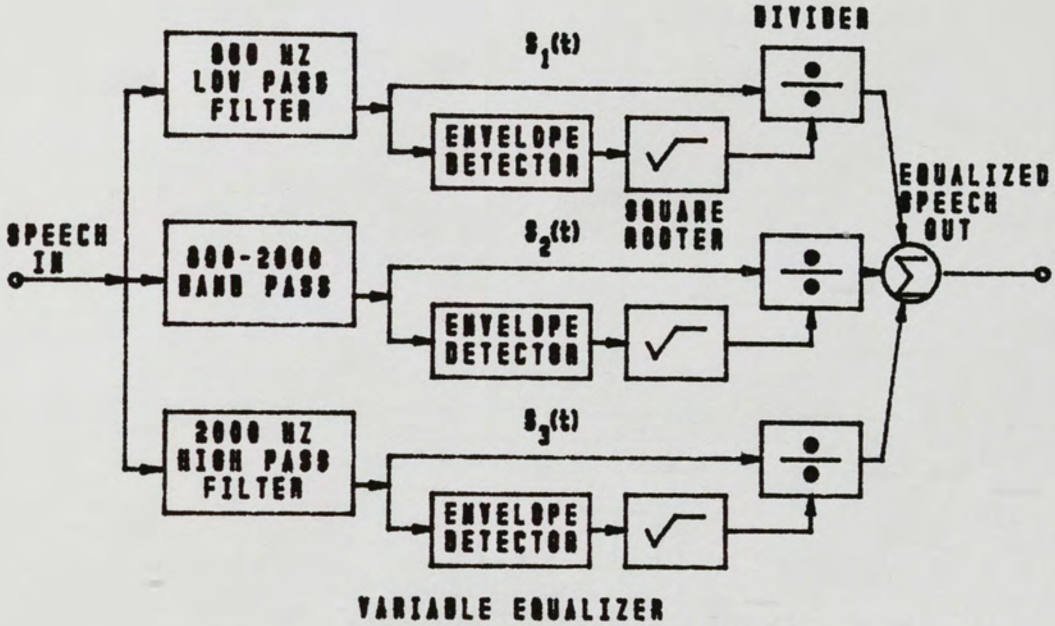


Figure 22. Block diagram of spectrum square rooter used in autocorrelation vocoder.

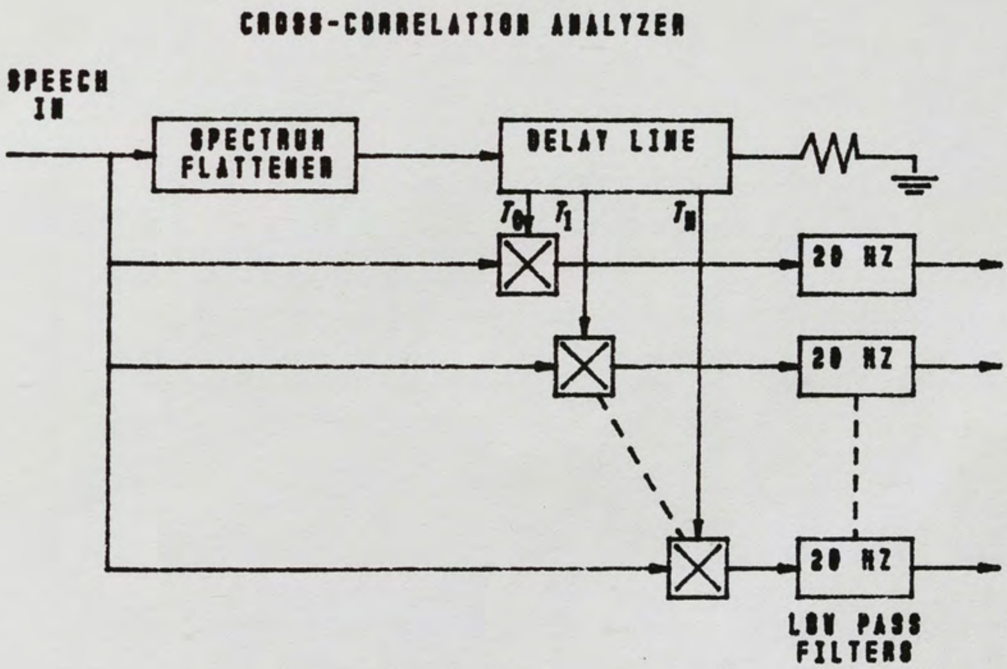


Figure 23. Block diagram of cross-correlation analyzer.

Voice-Excited Vocoder

Another type of vocoder is the voice-excited vocoder. The voice-excited vocoder generates the excitation signal required by the synthesizer from an uncoded sub-band (called the baseband) of the original signal [19]. This band is used to supply the excitation to the synthesizer. The wide-band excitation is processed from narrower baseband by nonlinear distortion techniques. This method produces either a flat spectrum of noise or harmonic frequency components [19]. Therefore, the excitation is produced from the original speech signal and not a coding process. A block diagram of a voice-excitation vocoder is shown in Figure 24 [19].

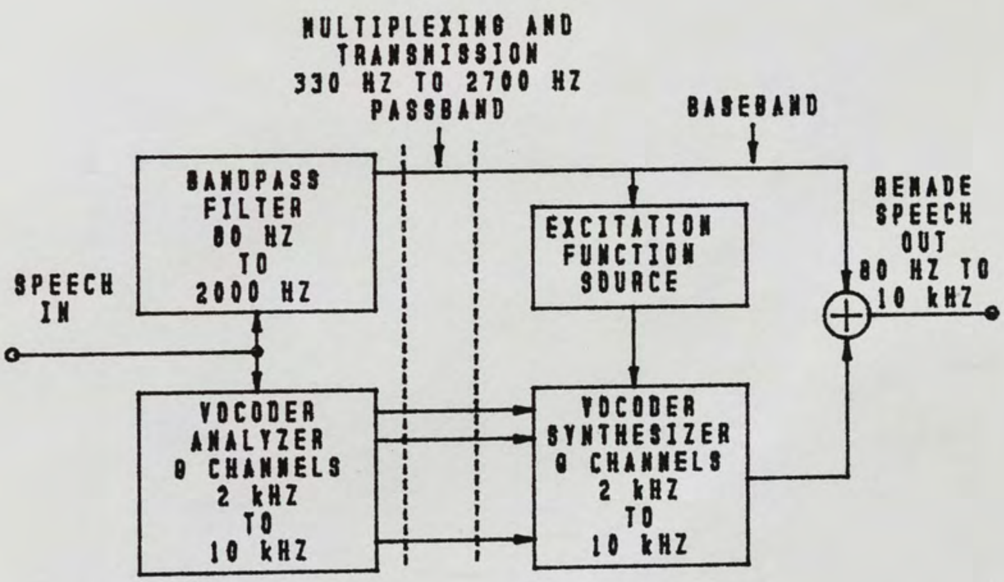


Figure 24. Voice-excited vocoder.

This type of vocoder eliminates pitch problems that are commonly related to vocoders. However, a problem does exist in that the system requires a larger transmission band because of the baseband.

Phase Vocoder

The phase vocoder specifies the speech signal in terms of its short-time amplitude and phase spectra. Like the voice-excited vocoder, the phase vocoder does not require a voiced-unvoiced or pitch detector as do channel vocoders.

In the phase vocoder, the original speech signal is input into a parallel bank of (n) band-pass filters and then recombined as the channel vocoders. The output of each filter can be represented as the convolution of the signal and the impulse response of the filter. Each convolution can be represented as a short-time Fourier transform of the original signal evaluated at some radian frequency. These transforms can be represented by their magnitudes, the short-time amplitude spectrums, and their phase, the short-time phase spectrums.

Unfortunately, the phase functions are generally unbounded and are not suitable as transmission parameters. However, the time derivatives of this phase function appear to be band-limited and are suited as transmission parameters [20].

The original signal can be reconstructed at the vocoder receiver/synthesizer. The phase functions are recovered by integrating the values of the derivatives. The output of a bank of (n) oscillators modulated in phase and amplitude are then summed. The oscillators are simultaneously phase and amplitude modulated from the band-limited versions of the amplitude and phase functions derived from the original speech signal [20].

The amplitude function of the phase vocoder is very similar to the conventional channel vocoder. The channel vocoder separates the excitation from the spectral envelope functions. The spectral envelope functions of the channel vocoder are the same as the amplitude functions of the phase vocoder. However, the excitation information is contained in a signal which determines pitch and voice-unvoiced signals. In the phase vocoder, the excitation is transmitted through the (n) phase functions [20].

Comparison of Different Schemes

Tables 4 [21] and 5 [21] are provided as general comparisons of different coding techniques. Table 4 compares the relative complexity of coders which is based on the relative count of logic gates. Table 5 compares the quality of coders and quality and the required bit rates.

TABLE 4

CODER COMPARISON

RELATIVE COMPLEXITY	CODER	
1	ADM:	adaptive delta modulator
1	ADPCM:	adaptive differentiAL PCM
5	SUB-BAND:	sub-band coder (with CCD filters)
5	P-P ADPCM:	pitch-predictive ADPCM
50	APC:	adaptive predictive coder
50	ATC:	adaptive transform coder
50	OV:	phase coder
50	VEV:	voice-excited vocoder
100	LPC:	linear-predictive coefficient (vocoder)
100	CV:	channel vocoder
200	ORTHOG:	LPC vocoder with orthogonalized coefficients
500	FORMANT:	formant vocoder
1000	ARTICULATORY:	vocal-tract synthesizer; synthesis from printed English text.

TABLE 5

QUALITY COMPARISON

QUALITY TRANSMISSION	CODER	kbits/s
TOLL-	Log PCM	56
	ADM	40
	ADPCM	32
	SUB-BAND	24
	Pitch Predictive ADPCM	24
	APC, ATC, OV, VEV	16
COMMUNICATIONS-	Log PCM	36
	ADM	24
	ADPCM	16
	SUB-BAND	9.6
	APC, ATC, OV, VEV	7.2
SYNTHETIC-	CV, LPC	2.4
	ORTHOG	1.2
	FORMANT	0.5

In addition to the above tables, other parameters which may be affected by the coder's application should also be compared. Some of these parameters include environmental tolerances, vibration tolerances, reliability. Other factors such as cost, availability, military or commercial standards, size and weight, should also be considered when selecting a coder. In general, the choice of a coder will always depend on its application.

CODED DATA ANALYSIS AND SYNTHESIS

Waveforms that are quantized or coded into a form of numbers can be stored as data. This data can then be analyzed, manipulated, digital filtered, modified, synthesized, etc., by speech processing techniques.

VAX computer programs were developed that utilize ILS speech processing programs to analyze a coded sample speech waveform [21]. The waveform was then synthesized by three different methods. The synthesized data was then compared to the original data. Listings of these programs appear in appendices A through D.

The first method analyzed the sample data using the ILS program (API), Analysis with Pitch Extraction Command. The API program analyzes the speech data using linear prediction and cepstral pitch detection.

The API command then models the spectral characteristics of the input data using autoregressive modeling. This modeling is accomplished by the linear prediction autocorrelation method (LPC). The method simultaneously determines the excitation (pitch period) and status periodic or non-periodic noise [21].

The ILS program (SNS), Synthesize Sample Data Command, was then used to synthesize the sample data. The

SNS command uses the analyzed parameters to synthesize the waveform with pitch synchronous interpolation or frame by frame.

Figure 25 illustrates the results of this procedure. Figure 25a is the sample waveform and 25b is the synthesized waveform. The following speech sample used to generate these speech signals: "Hello. How are you today"? The speech sample used less than 300 frames and lasted less than 2 seconds. Note that the signals are not identical. Error and distortion occurred when the signal was reproduced from the parameters which were obtained from the original signal analyzsis. The program used for this analysis and synthesis is included as Appendix A.

Even though all the parameters of both signals are identical, an error between the two signals does exist. This error is due to the constraints of the system and error introduced during the analysis and synthesis process.

To illustrate this error the ILS program (VTR), Vocal Tract Command [21], was implemented. The VTR program plots a reproduction of the midsagittal vocal tract area function using the coefficients stored in the analysis file. Each vocal tract image represents one analysis frame. Figure 26a illustrates the sample signal and 26b the synthesized signal.

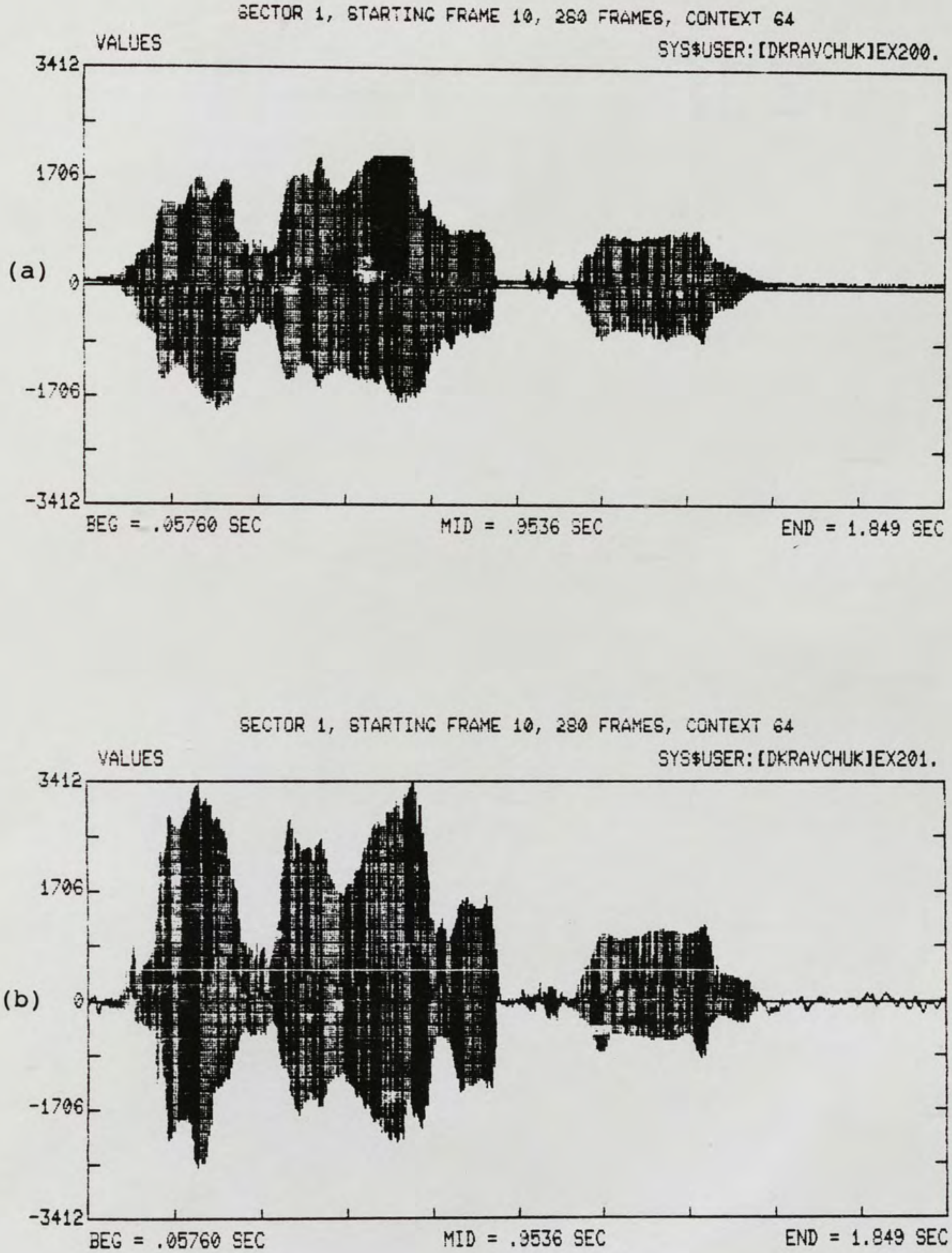


Figure 25. API Analysis and SNS Synthesizer.
(a). Sampled Data
(b). Synthesized Data

(a)

(b)

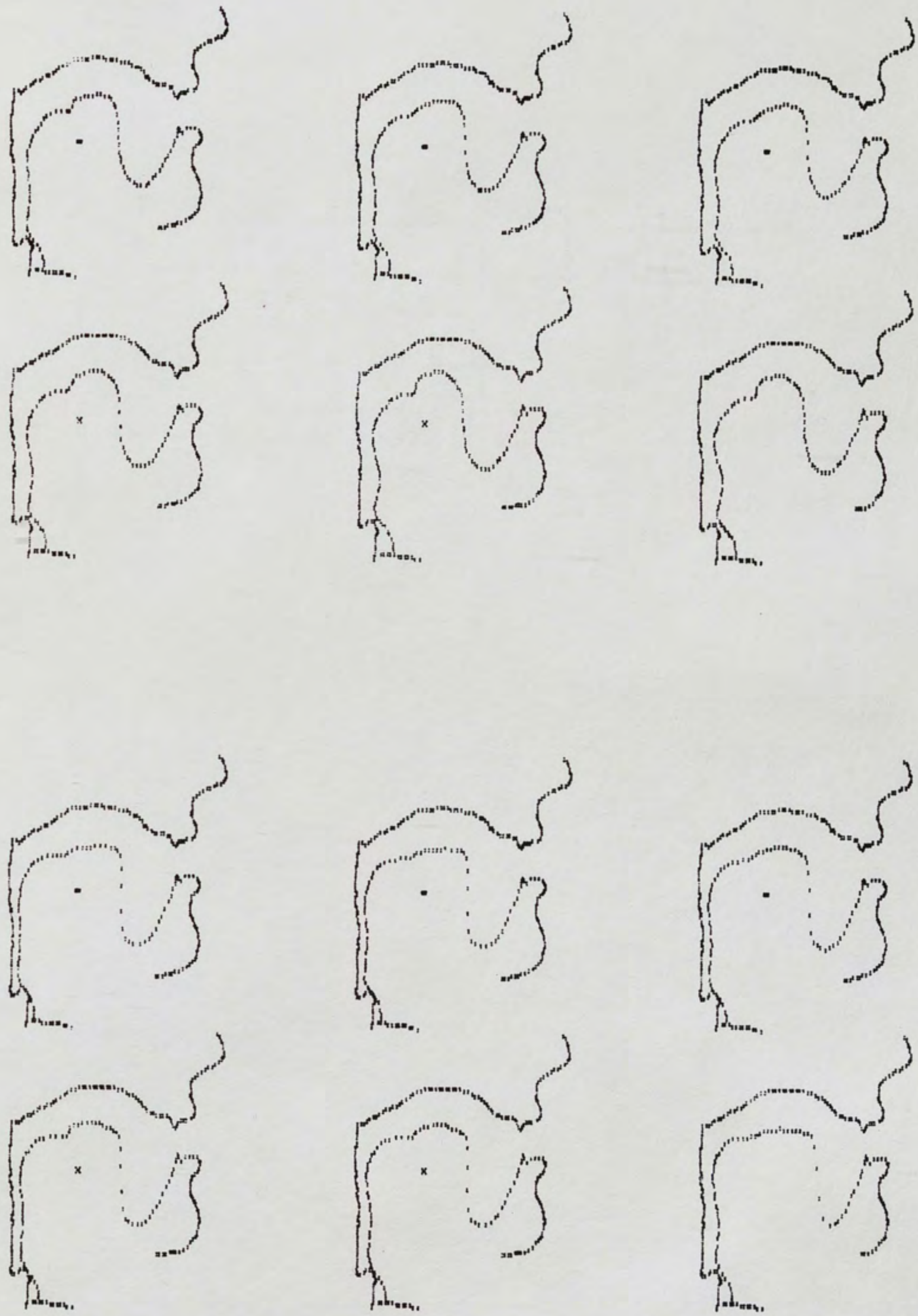


Figure 26. Vocal Tract Reproduction of API Analysis and SNS Synthesize
(a). Sampled Data
(b). Synthesized Data

The second method of analysis and synthesis uses the ILS programs PAN and PNS. The (PAN), Pitch Analysis Command, performed autoregressive analysis on the sample signal [21]. Unlike the API command which starts at the beginning of each data frame and uses a fixed analysis window size, the PAN command allows for variable starting points and window sizes. The starting point of each analysis window is labeled by the ILS program (CLA), Labeled Sampled Data Segments Command [21]. The labels mark the beginning and end of voiced and unvoiced sections of speech. Each pitch period is labeled for voiced sections and the entire section is labeled for unvoiced section. The labeled unvoiced section is automatically divided into 10 msec segments for analysis.

The analysis program (PNS), Pitch Synchronous Synthesis Command was then used to synthesize the analyzed data. The PNS command uses the analyzed parameters to generate speech pitch synchronously. The program matches the pitch and gain locations to those of the sampled waveform. Figure 27 illustrates the results of this procedure. Figure 27a is the sample signal and 27b is the synthesized signal.

To compare the two waveforms, the synthesized signal was analyzed using the API command and compared to the API analysis of the sample signal. As in the first method, all the parameters are identical. The program used for this analysis and synthesis is included as Appendix B.

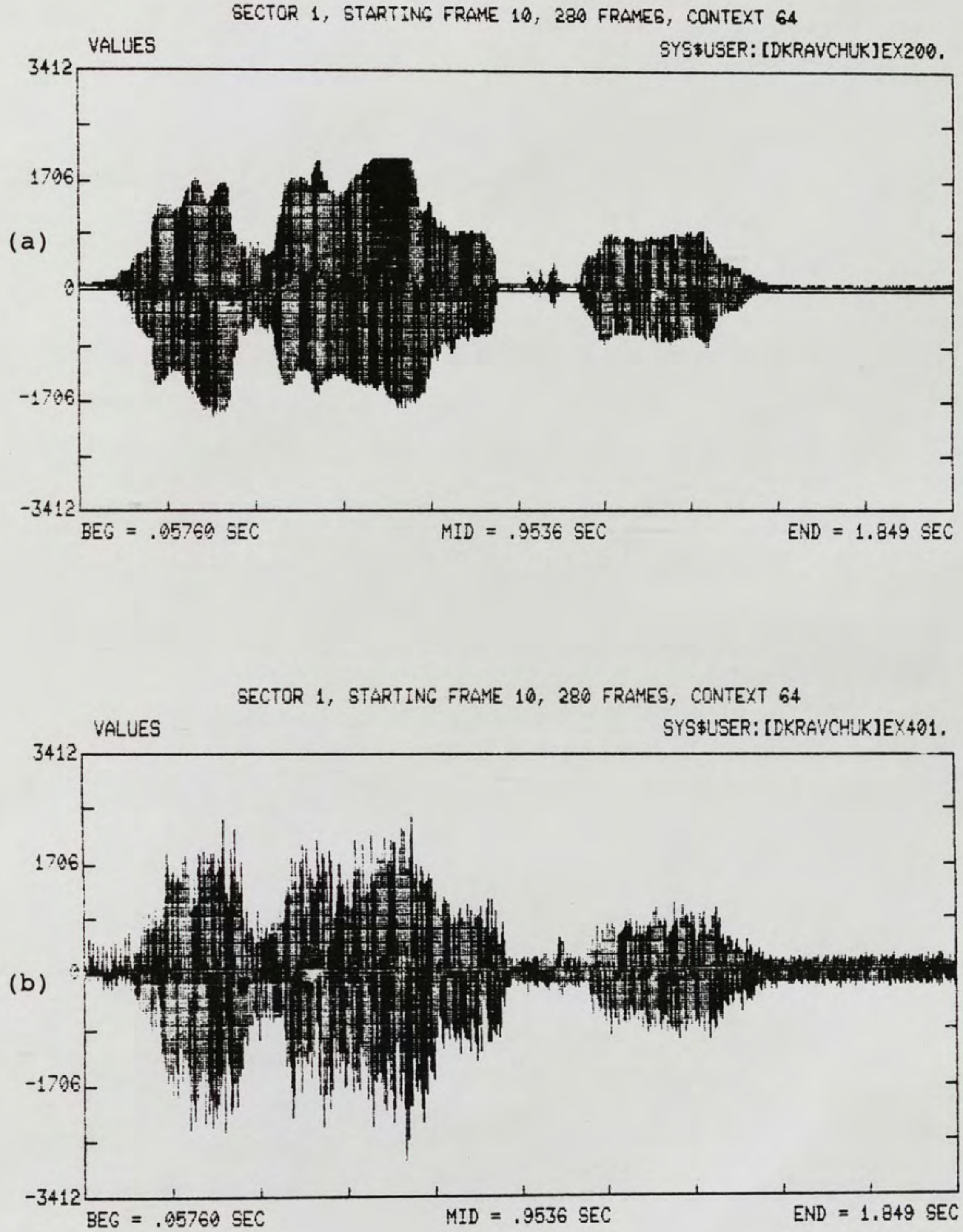


Figure 27. PAN Analysis and PNS Synthesize.
(a). Sampled Data
(b). Synthesized Data

Figure 28 illustrates the error between the two signals by using the VTR command.

The third method of analysis and synthesis uses the ILS programs ANA and SNS. The (ANA), Analysis Command is similar to the API command. The ANA performs linear predictive (autoregressive) analysis. The analysis is based on a sample signal being approximated by a linear combination of past samples [21]. The Synthesize Sample Data (SNS) command is the same used in the first method.

Figure 29 illustrates the results of this procedure. Figure 29a is the sample signal and 29b the synthesized signal. Error and distortion are present in the synthesized signal.

The synthesized signal was again analyzed using the API command and compared to the API analysis of the sample signal. Again, the parameters are identical. The program used for the analysis, synthesis and comparison is enclosed as Appendix C.

Again, as in the first two methods, the error between the two signals can be seen using the VTR command in Figure 30. A copy of the VTR program used for all three methods is included as Appendix D.

(a)

(b)

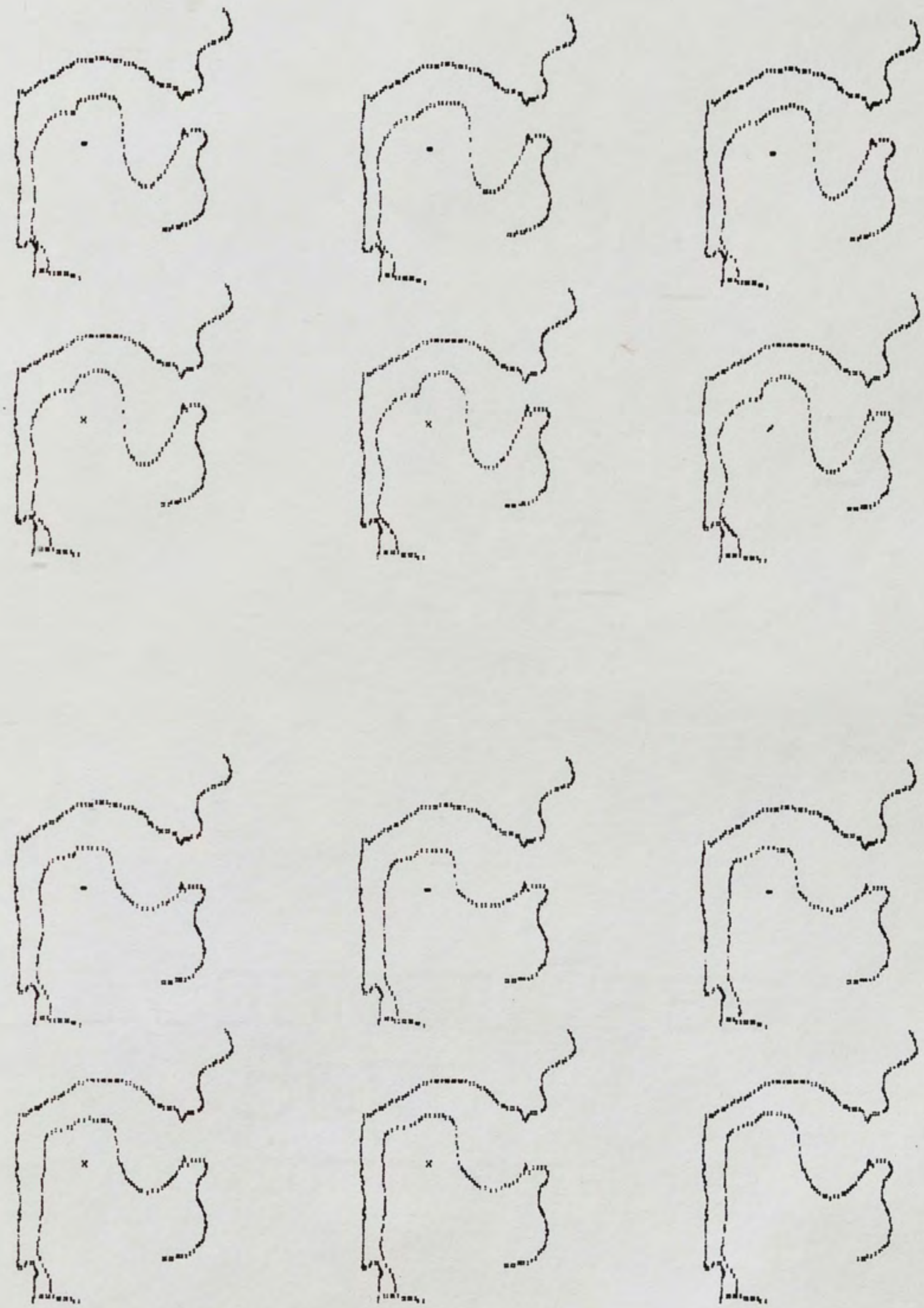
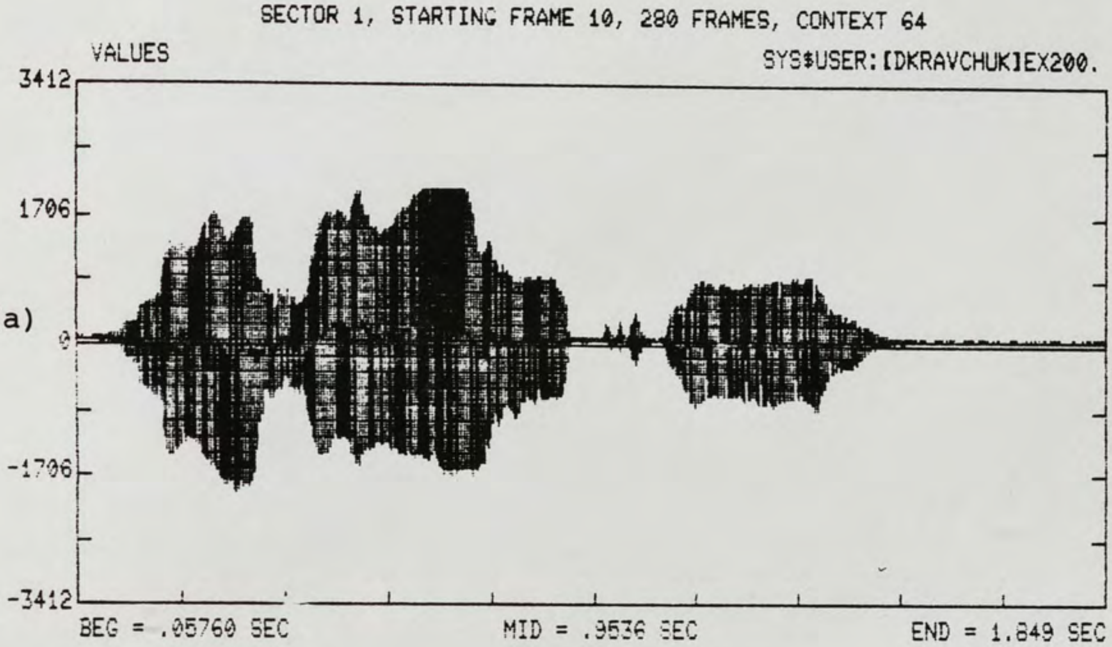
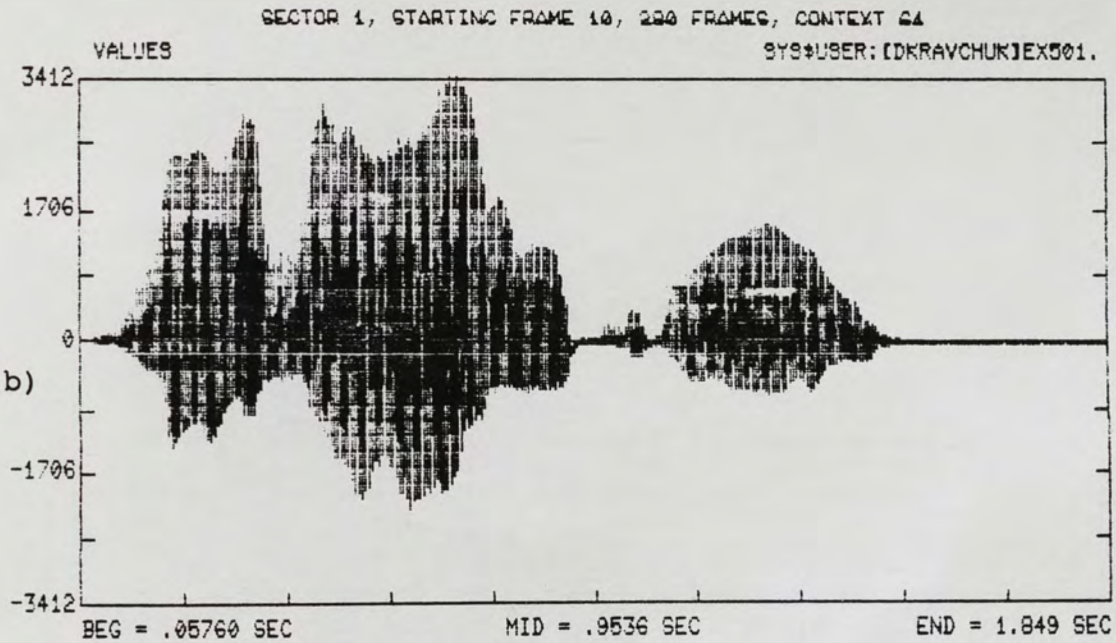


Figure 28. Vocal Tract Reproduction of PAN Analysis and PNS Synthesize.
(a). Sampled Data
(b). Synthesized Data



(a)



(b)

Figure 29. ANA Analysis and SNS Synthesize.
(a). Sampled Data
(b). Synthesized Data

(a)

(b)

Figure 30. Vocal Tract Reproduction of ANA Analysis and SNS Synthesize.
(a). Sampled Data
(b). Synthesized Data

APPENDIX A

```
!DSP2.COM [SNS,API]
$FIL ANEX
$FIL DEY201,,6
$FIL 200
$FIL S202
$API 10,280
$FIL S
$FIL B201
$SNS B10,280
$FIL 201
$FIL S203
$API 10,280
$FIL S
$FIL 200
$DSP E10,280,60
$FIL 201
$DSP E10,280,60
[EOB]
```

APPENDIX B

```
!DSP4.COM [SYN FROM PITCH SYN ANAL DATA (PAN,PNS)]
$FIL DEY401,,5
$FIL 200
$LBF DEYTEST1
$TSI L
$FIL S402
$PAN 1,-1
$FIL S
$FIL B401
$PNS B1,-1
$FIL 401
$FIL S403
$API 10,280
$FIL S
$FIL 200
$DSP E10,280,60
$FIL 401
$DSP 10,280,60
[EOB]
```

APPENDIX C

```
!DSP5.COM [SYN USING LINEAR PREDICTION (ANA,SNS)]
$FIL DEY501,,6
$FIL 200
$FIL S502
$ANA 10,280
$FIL S
$FIL B501
$SNS B10,280
$FIL 501
$FIL S503
$API 10,280
$FIL S
$FIL 200
$DSP E10,280,60
$FIL 501
$DSP E10,280,60
[EOB]
```

APPENDIX D

```
!VOCAL TRACT PRINT
$FIL DEY700,,6
$FIL 200
$FIL S700
$ANA 50,15
$FIL S700
$VTR 50,6
$FIL 201
$FIL S702
$ANA 50,6
$FIL S702
$VTR 50,6
$FIL 401
$FIL S704
$ANA 50,15
$FIL S704
$VTR 50,6
$FIL 501
$FIL S705
$ANA 50,15
$FIL S705
$VTR 50,6
[EOB]
```

BIBLIOGRAPHY

[1] Oppenheim, Alan V., and Schafer, Ronald W. Digital Signal Processing. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1975.

[2] Rabiner, L. R., and Schafer, R. W. Digital Processing of Speech Signals. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1978.

[3] Lathi, B. P. Modern Digital and Analog Communication Systems. New York: CBS College Publishing, 1983.

[4] Mangold H. Analysis, Synthesis, and Transmission of Speech Signal, Advisory Group for Aerospace Research and Development Lecture Series No. 129. Loughton Essex, Germany: Specialized Printing Services Limited, 1983.

[5] Schafer, R. W., and Rabiner, L. R. Digital Representation of Speech Signals. IEEE Transaction: Speech Analysis. New York: IEEE Press, 1979.

[6] Lathi, B. P. Communication Systems. New York: John Wiley & Sons, Inc., 1968.

[7] Flanagan, J. M. Schroeder, Atal B., Crochiere, R., Jayant, N., and Tribolt, J. Speech Coding, IEEE Transactions on Communications. New York: IEEE Press, 1979.

[8] Cattermole, K. W. Principles of Pulse Code Modulation. New York: American Elsevier Publishing Company, Inc.,1969.

[9] Schwartz, Mischa. Information Transmission, Modulation, and Noise. New York: McGraw-Hill Book Company, 1959.

[10] Baghdady, Elie J., ed. Lectures on Communication System Theory. New York: McGraw-Hill Book Company, 1961.

[11] Jayant, N. S. Digital Coding of Speech Waveforms: PCM, DPCM and DM Quantizers. IEEE Transaction: Waveform Quantization and Coding. New York: IEEE Press, 1976.

[12] Schwartz, Mischa, and Shaw, Leonard. Signal Processing: Discrete Spectral Analysis, Detection, and Estimation. New York: McGraw-Hill Book Company, 1975.

[13] Tribolet, Jose, M., and Crochiere, Ronald, E. Frequency Domain Coding of Speech. IEEE Transactions on Acoustics, Speech, and Signal Processing. New York: IEEE Press, 1979.

[14] Gold, B., and Rader, C.M. Digital Processing of Signals, New York: McGraw-Hill, 1969.

[15] Hancock, John C. An Introduction to the Principles of Communication Theory. New York: McGraw-Hill Book Company, Inc., 1961.

[16] Rabiner, Robert R., and Gold, Bernard. Theory and Application of Digital Signal Processing, Englewood Cliffs: Prentice-Hall, Inc., 1975

[17] Markel, J. D., and Gray, H. Jr. Linear Predication of Speech Signals, Berlin Heidelberg, New York: Springer-Verlag, 1976.

[18] Schroeder, M. R. Vocoders: Analysis and Synthesis of Speech. IEEE Transaction: Speech Analysis. New York: IEEE Press, 1979.

[19] David, E., Jr., Schroeder, M. R., Logan, B. F., and Prestigiacomo, A. J. Voice-Excited Vocoders For Practical Speech Bandwidth Reduction. IEEE Transactions: Speech Analysis. New York: IEEE Press, 1979.

[20] Flanagan, J. L., and Golden, R. M. Phase Vocoder. IEEE Transactions: Speech Analysis. New York: IEEE Press, 1979.

[21] Signal Technology Inc. Interactive Laboratory System Software Manual. 1985.