EFFICIENT CODING OF THE PREDICTION RESIDUAL

By

LEGAND L. BURGE, JR.

Bachelor of Science
Oklahoma State University
Stillwater, Oklahoma
1972

Master of Science
Oklahoma State University
Stillwater, Oklahoma
1973

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the Degree of
DOCTOR OF PHILOSOPHY
December, 1979

EFFICIENT CODING OF THE PREDICTION RESIDUAL

Thesis Approved:

Thesis Adviser

Dean of the Graduate College

ACKNOWLEDGMENTS

I extend my sincere thanks and gratitude to Dr. Rao Yarlagadda, my thesis adviser and chairman of my doctoral committee, for his dedication and guidance during the term of my graduate program. His interest and encouragement through understanding has contributed significantly to realize the completion of this dissertation. My appreciation is due to Drs. Bennett Basore, Craig S. Sims, and Ronald D. Schaefer, my committee members, for their helpful comments and generous contribution of time throughout this research.

I owe special recognition to Dr. Cheryl Scott for stimulating discussions concerning aspects of speech science.

A special thanks is due to Drs. Charles Bacon, H. Jack Allison and Robert Mulholland for stimulating discussions and encouragement during this work. A sincere appreciation to Dr. Lynn R. Ebbesen and John Perrault for their assistance and accessibility of computer time.

I would like to thank the United States Air Force for providing me the opportunity to complete my degree through the AFIT Civilian Institutions Program. I am grateful to Lt. Col. John Kitch, Jr., and Capt. Samuel Brown, Jr., for their understanding and helpfulness during my academic tenure.

I am grateful to the Defense Communications Engineering Center (DCA) in Reston, Virginia, for their support during the term of the research. The stimulating discussions and kindness awarded me by Mr. Gerald Helm

and Mr. George Moran is appreciated. The assistance of Mr. William Mills and Ms. Elaine Bernd was greatly appreciated during the term.

My sincere gratitude to Ms. Debbie Perrault and Ms. Louise Sumpter for their excellent typing of this dissertation. I would like to acknowledge Mr. Eldon Hardy for the excellent drawing of the figures.

I owe special thanks to Mr. Legand L. Burge, Sr., Mrs. Bobbie J. Burge, Mr. Richard Jones, Mrs. Wilba Jones, Mrs. Annie Dean, Rev. Dr. and Mrs. Joe Edwards, Rev. and Mrs. Richard Thompson, Mr. and Mrs. Glenn Mosely, Mr. Lynn R. Osborn, Mr. Lew Phillips, and Mr. Raymond A. Young for their encouragement throughout this work.

The cooperation, patience and understanding of my wife, Claudette, has been greatly appreciated. My son and daughter, Legand and LeAnn, are acknowledged for the continued source of motivation for which this effort was begun.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

M	-	Total number of data points
s_n	-	Signal
a_i	-	Forward lattice LPC filter coefficients
b_i	-	Backward lattice LPC filter coefficients
$R(k)$	-	Residual signal $k = 0, 1, \ldots, M-1$
p	-	Order of the LPC filter
$F_1, F_2, \ldots$	-	Formant frequencies
$r_\ell(m)$	-	Output of the ℓth bandpass filter $m = 0, 1, \ldots, M-1$; $\ell = 1, 2, \ldots, B$
$r_e(k)$	-	Decimated signal $r(k)$
ω_ℓ	-	Bandpass filters are assumed to be contiguous; ℓ gives the lowest frequency corresponding to the ℓth bandpass filter
$h_n(k)$	-	Impulse response of nth lowpass filter associated with the nth sub-band
W_ℓ	-	Bandwidth of the ℓth bandpass filter
ω_n	-	Edges of the bandpass filter
$e_n(k)$	-	Modulated signal of sub-band
N, N_i	-	Samples available per frame
f_o	-	Fundamental frequency
$f_1, f_2, \ldots$	-	Harmonics of the fundamental
f_s	-	Original sampling frequency
f_{si}	-	Sampling frequency for the ith sub-band

$g(t)$, $G(\omega)$	-	Represents glottis excitation function in the time and frequency domains
$v(t)$, $V(\omega)$	-	Represents vocal tract filter in the time and frequency domains
$s(t)$, $S(\omega)$	-	Represents speech waveform in time and frequency domains
$x(n)$, $y(n)$, $x(k)$, $y(k)$	-	Speech sample sequency
$x_f(n)$	-	Forward prediction of $x(n)$
$x_b(n)$	-	Backward prediction of $x(n)$
$e_f(m)$	-	Forward prediction error or prediction residual signal
$e_{fn}(k)$	-	Output of the nth bandpass filter
$e_b(k)$	-	Backward prediction error
R_i	-	Autocorrelation of $x(n)$
E_j	-	Residual signal at the jth stage
C_j	-	Cross correlation between forward and backward prediction errors
k_{j+1}	-	Partial correlation coefficients (PARCOR) of (j+1)st stage
$A_p(z)$	-	Transfer function between forward prediction error and speech sequence for pth stage
$B_p(z)$	-	Transfer function between backward prediction error and speech sequence for pth stage
r_{j+1}	-	Reflection coefficient of (j+1)st stage
σ^2	-	Variance of speech source
y	-	Transform coding coefficients
A	-	n x n matrix

$\overline{D}$	-	Mean-square overall distortion
J_i	-	Number of bits/sample
δ	-	Correction factor for a quantizer
$\overline{R}$	-	Average bit rate
R_{xx}, R_{yy}	-	Covariance matrices for x and y
λ_i	-	Eigenvalue of R_{xx}
k_{ij}	-	Bits/sample corresponding to the ith energy and jth frequency band
$E_{fn}(k)$	-	Discrete Fourier transform coefficient
E_n	-	Energy of the nth sub-band
E_{ij}	-	Energy of the signal for the ith energy band and jth frequency band
E^T_{ij}	-	Energy threshold for the ith energy band and jth frequency band
k^A_{ij}	-	A priori bits/sample corresponding to the ith energy band and jth frequency band
σ_i	-	Normalization factor
α_n	-	Discrete Cosine Transform coefficient
Δ	-	Quantizer step size
ψ	-	Quantizer array
Q_e	-	Quantization value array
ζ_ℓ	-	Quantizer level

CHAPTER I

INTRODUCTION

1.1 Statement of the Problem

The structural unit of speech composition is the speech sound called the phoneme.[†] Its variations are called allophones. It can be said also that phonemes relate to the linguistic basis of a language. However, phonemes are not "bricks," i.e., the human has been endowed with the ability to communicate in a continuous mode. Because we speak in an uninterrupted fashion in order to complete our thoughts, the phonemic structure connects itself by transitional cues for the perception of certain phonemes [1]. It is this transitional information that is needed for absolute discrimination of speech and speech-like sounds [2]. It is the transitional information that is needed for efficient excitation of a speech synthesizer.

To synthesize intelligible speech, the perceptual aspects of speech sounds have to be used. In other words, the ability for humans to discriminate and differentiate a speech sound with their over-learned senses must be incorporated into the speech synthesis technique. The speech synthesis must include perceptual enhancement, and the inclusion of transitional information (that is, frequency shifts). Transitional information is the loci of frequency determined by the place of

---

[†]Some of the words related to the science of the speech waveform are defined in APPENDIX A.

articulation that connects the phonemes. Phonemes are the basic speech sound element used to make a word. One can also say that a phoneme is an idealized structural unit of language which serves to keep words apart. It is an astonishing fact as to how the human brain stores rules to keep track to one's language for communicating. The object of speech synthesis is to come as close as possible to this occurrence.

The history of synthetic voice coding had its origination with H. W. Dudly in 1939 [3] [4]. The Dudley speech reproduction model consists of a filter representing the vocal tract resonance characteristics driven by an artificially synthesized excitation signal. The filter and the excitation signal parameters are updated periodically. To determine the filter characteristics, Dudley used the Fourier spectrum of the speech as a basis. The excitation signal consists of a pulse train for voiced sounds and random noise for unvoiced sounds. The model that Dudley has represented is essentially the basis of many methods today [5] [6] [7]. Some of these ideas are discussed below.

A basic model of the speech waveform is to assume a linear quasi time-invariant system which responds to a periodic or noiselike excitation. This linear time invariant system represents the vocal tract. If the vocal tract is assumed to be fixed, then the output of the system is a convolution between the excitation and vocal tract transfer function (see Figure 1).

Recently considerable interest has been given to methods of digital analysis and synthesis of speech assuming the presented model. A method that has proven to be efficient for encoding the speechwave is linear prediction [6]. The linear predictive encoder was developed to improve the channel vocoder voice quality and intelligibility [7]. The difference

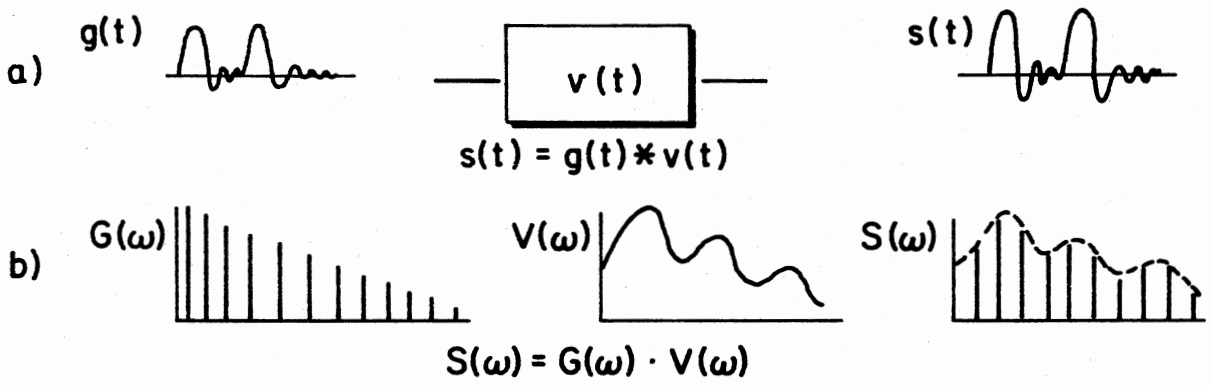


Figure 1. Model of Speech Production as the Response of A Quasi-Stationary Linear System (a) Time Domain Characterization, (b) Frequency-Domain Characterization (After Oppenheim, 1978)

between the linear predictive coded (LPC) vocoder and the channel vocoder is the filter. There are two types of LPC vocoders, a pitch-excited and a residual-excited. The difference between the two is how the excitation signal is characterized for the synthesis filter. In the pitch-excited LPC vocoder, the model of the vocal tract, with glottal flow and radiation, is represented by the predictor coefficients. These coefficients are transmitted together with the information regarding the excitation of the speech, i.e., pitch, voiced/unvoiced decision and the gain. Much research has been done toward the pitch-excited LPC vocoder. Two methods have been discovered, the autocorrelation [8] [9] and the covariance [6] methods. The residual-excited methods can be characterized the same way. However, instead of using pitch, voiced/unvoiced desicion and gain, the residual is encoded and transmitted. The residual is the difference between the actual and predicted speech signals. This technique also carries the name adaptive predictive coding (APC). The channel vocoder, on the other hand, uses a set of narrowband filters whereas the linear predictor uses an all pole digital filter. The linear predictive filter describes the frequency response of the vocal tract system by the predictor coefficients. Its function is to decompose the speech into two waveforms. One waveform represents the parameters that are time-varying such as predictor coefficients, partial correlation coefficients and other parameters that represent the formant frequency characteristics. The other waveform is the prediction residual. Figure 2 describes a block diagram of the LPC analysis.

The prediction residual is the ideal signal for an excitation function for the linear predictive analysis and synthesis model because it contains the actual information instead of the pseudo-model, a pulse

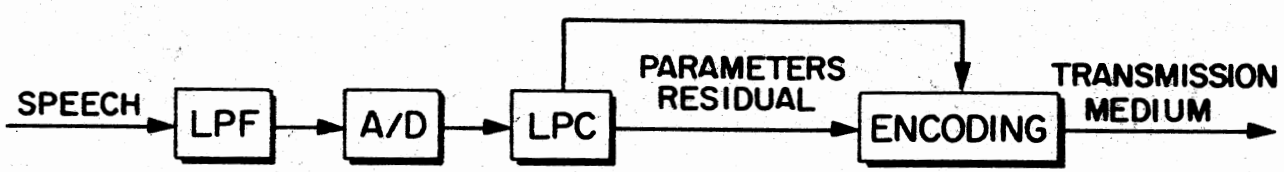


Figure 2. Block Diagram of LPC Analysis

train or random noise [10]. In addition, phasing information is embedded in the prediction residual. Furthermore, since the analysis filter is the inverse of the synthesis filter, the decomposed waveform can be reconstructed to form the input speech waveform by updating the parameters [7] [11]. The prediction residual also follows the actual speech excitation model g(t) in Figure 1 [12]. The function, g(t), represents the glottal pulse which is also called the glottal volume velocity at the vocal cords or glottis. In order to ideally model the voice reproduction system, it is necessary to use a system whose properties are similar acoustically to the glottis and vocal tract. It is best to model the excitation signal with an analogous function to the glottis waveform for input to the vocal tract. It is well known that for nonnasal voiced speech sounds, the transfer functions have no zeros [5]. For these particular sounds, the vocal tract filters can be approximated by an all pole filter. It is also known that the shape and periodicity of the glottis excitation are subject to large variations [12]. However, with the linear predictive model the features of the glottal flow, the vocal tract, and the radiation, which is the output from the mouth, are included into a single recursive filter. To separate the glottal flow from the vocal tract involves a deconvolution. Some authors have avoided this separation of the source function; however, the artificial excitation used by them represents only a good approximation to the prediction residual for unvoiced sounds. Moreover, for voiced sounds, the artificial excitation could be improved. The prediction residual should be used for the excitation function, because it contains the following characteristics:

1. It is repetitive at the pitch frequency.
2. It has basically a flat amplitude spectrum; however, it includes

details that relate to the suprasegmentals of the individual and of the spoken words.

3. It includes the noisiness of opening and closing of the glottal mechanism indicating phase information.

4. It includes the fact that voiced fricatives and stops are a combination of noise and a repetitive signal.

Noting the speech characteristics in the residual signal, several authors have investigated the coding aspects of the prediction residual [13-32]. However, the speech intelligibility aspects, such as Articulation Index (AI) [29], have not been used in these. The Articulation Index concept has been used effectively in the sub-band coding of speech [36]. The sub-band coding, based upon AI, allows for an efficient bit distribution in coding. This thesis combines all these ideas and presents an efficient method of coding the prediction residual using the concepts of sub-band coding. A literature survey related to these areas is presented in the next section.

One important aspect of coding is bit rate. For certain narrow band rates, the coding of the prediction residual is not feasible [13]. Also, it has been shown that 9,600 bits/second is feasible for transmission of residual and filter parameters, and is practical over voice grade lines [35]. In the future, lower data rates have to be used for cost effectiveness. At present, rates below 6,000 bits/second yield speech quality of a synthetic nature. Rates between 6,000 bits/second and 16,000 bits/second demonstrates good communication quality. Studies have shown and present operating equipment demonstrate that a 16,000 bits/second transmission rate and above yield toll telephone quality. The thrust of the governmental community for designing voice switch networks has been

recently toward 9,600 bits/second rate. At this rate, the communicators can comprehend the language spoken; however, there is some drop-off in speaker recognition but not as drastic as at rates closer to 6,000 bits/second. With the advent of mircroprocessing systems more sophisticated algorithms can be implemented with small monetary investments. This thesis presents the coding and decoding of the residual signal using sub-band coding at a data rate of 9,600 bits/second.

1.2 Review of the Literature

Predictive systems related to speech have evolved through the years. A brief survey of these systems is presented below. In earlier studies of predictive coding systems with applications to speech signals, the linear predictors were limited to fixed coefficients in an interval [17]. In more recent studies, it was found that since the speech signal has non-stationary properties, the linear predictor does not efficiently predict the signal at each interval. In work by Atal and Schroeder [6], an adaptive predictive system took into account the quasi-periodicity of speech signals. In addition to being the classic forerunner for adaptive predictive coding (APC) of speech signals, this is a more elaborate predictor than the one with fixed coefficients which is suited for characteristics of speech sounds. Basically, the residual signal along with the predictor provides sufficient information for the receiver to regenerate the input. In this, pitch is determined from the residual signal. Atal and Schroeder [22] have examined predictive coding of speech signals recently. They have shown that speech quality can be improved by masking quantizer noise over the speech signal. Atal and Hanauer [5] described an efficient encoding of the speech wave by representing it in terms of time-varying

parameters related to a transfer function of the vocal tract and by modeling the excitation.

In work by Dunn [13], the linear predictive coded residual signal was generated by a feed-forward linear predictive coding (LPC) analyzer and encoded using delta modulation (DM). The signal was transmitted at a bit rate of 9,600 bits/second. Gibson, Jones, and Melsa [14] have introduced a method called sequential adaptive prediction which utilized differential pulse code modulation (DPCM) with an adaptive quantizer and an adaptive predictor using Kalman filtering. This work was improved upon by Cohn and Melsa [15] using adaptive differential pulse code modulation (ADPCM) for encoding the prediction residual. A method using the Kalman filter for the adaptive predictive encoder was introduced by Goldberg and others [16]. This system was real time APC that was implemented on a minicomputer. An adaptive residual coding using an adaptive predictor, adaptive quantizer, and a variable length coder was studied by Qureshi and Forney [18]. In these studies, a class of speech digitization algorithms is described for use at bit rates of 9,600 to 16,000 bits/second. These systems involve an adaptive predictor, an adaptive quantizer, and a variable length coder. This is a practical version of a residual encoder previously studied by Melsa and others [14]. Most recently, the method of variable length coding of the prediction residual was studied by Berouti and Makhoul [19]. This system of APC uses a noise spectral shaping filter to solve the granular noise quantization problem and an indefinite quantizer to solve the overload quantizing problem.

A voice-excited predictive coder (VEPC) by Esteban and others [20] uses a baseband excitation of the residual and splitband coding by signal decimation/interpolation. Furthermore, quadrature mirror filters are

implemented in order that the aliasing properties could be taken advantage of in the synthesizer.

The most recent work by Cohn and Melsa [21] [23] involves the implementation of a speech coding algorithm for digital transmission of speech at 9,600 bits/second using a sequential, adaptive linear predictive coder, an adaptive source coder, and multipath tree-searching algorithm to generate quality speech. This is an extension of the previous work done on a residual encoder which was an improved ADPCM system for speech digitization. Chang [24] has extended this work and incorporated a noise resistant code for transmission.

In work by Magill and others [25], a feed-forward LPC analyzer was used with an encoding method of Adaptive Delta Modulation (ADM) and an experimental method of encoding the residual by DPCM. This is referred to as a residual excited linear predictive (RELP) vocoder. It combines the advantages of linear predictive coding and voice-excited vocoding.

Recently, Dankberg and Wong [26] have implemented a new version of the RELP vocoder. Their results have included a development of a pitch predicted ADPCM residual encoder and a harmonic generator. Viswanathan and others [27] considered the use of voice-excited linear predictive (VELP) and RELP coders for speech. They have studied in detail the various aspects of these coders and have attempted to maximize speech quality as a result. They also studied the advantages and disadvantages of baseband residual transmission and baseband speech transmission.

In recent work, Kang [28] studied the development of a narrowband voice digitizer that improves speech quality, intelligibility and reliability. The principle of LPC is used in implementing the lattice filter for the analysis and synthesis. Itakura and Saito [9] [30] have used

the lattice method for LPC analysis of speech. The thrust has been for improved quantization of partial correlation (PARCOR) coefficients. Makhoul [31] has presented a class of stable and efficient lattice methods for linear prediction of speech. In this, an indepth study is made on PARCOR coefficients. If the all pole function is stable, then the lattice obtained from this is stable; furthermore, since the PARCOR coefficients are bounded, stability is guaranteed and an efficient quantization method can be used.

In work by Flanagan [32], it is shown that the residual approximates the glottal waveform. In any excitation system, the closer one can approximate the physical model, the better response one gets from the system. Flanagan's work enhances this concept to use the residual waveform as the excitation to the speech synthesizer.

Rabiner and others [33] have studied the LPC error signal. The work investigated the variation of the prediction error as a function of position in an analysis frame within a single stationary speech segment. The error signal has the frequency range of the actual speech.

The work of Goodman [34] found the analog signal can be divided into several nonoverlapping frequency bands. Each band can be sampled and quantized independently. The result is an improvement in encoding efficiency over straight sampling and quantizing of signals that are spectrum peaked. Crochiere and others [36] [37] have applied this to speech signals in the digital domain. This is referred to as sub-band coding (SBC). This approach provides a means of controlling and reducing quantization noise in the coding.

A pilot study of speech waveform coding techniques were studied by Tribolet and others [38]. The study compared subjective ratings to the

various quality (objective) measures for speech waveform coders. Tribolet and others examined four different speech waveform coder algorithms for low-bit rate applications, and studied these relationships for overall objective and subjective ratings for quality. The algorithms were: adaptive differential PCM with a fixed predictor (ADPCM-F), sub-band coding (SBC), ADPCM with a variable predictor (ADPCM-V) and adaptive transform coding (ATC). The transmission rates studied were 24,000, 16,000, and 9,600 bits/second. The objective measures used were a conventional signal-to-noise ratio, frequency weighted signal-to-noise ratio, log likelihood ratio, and an articulatory bandwidth measure. The results of the study were that if complexity/cost was of no concern, then ATC is the most attractive of the group coders. However, if complexity/cost was a concern, then SBC is an attractive choice. ADPCM-F had the poorest quality for its complexity; ADPCM-V was the most costly for its quality. The transform coding and the sub-band coding will be explained in detail in Chapter II.

In the work by Barabell and Crochiere [39] a new design of the sub-band coding has been implemented for low-bit rate coding of speech. This study applied quadrature filters to SBC. This method has also employed pitch prediction within the sub-bands. Crochiere [40] has implemented a novel approach for pitch extraction in the SBC. The method uses digital linear phase shifters based on a bandpass interpolation scheme to achieve the non-integer delays necessary in the feedback loop for the pitch predictors. It uses the fractional sample delay in the pitch loop and permits the processing of the pitch prediction in each sub-band to be performed at the sub-band sampling rate which contributes to the efficiency of the algorithm.

Pitch detection algorithms that have been mentioned above have one basic goal. That is, make a voiced or unvoiced decision and during certain periods of voiced sounds, estimate the pitch period.

There are three areas of categorization for pitch detectors. First, there is a group that uses time-domain properties of speech signals. These pitch detectors operate directly on the speech waveform in order to estimate the pitch period. The measurements that are usually taken are minimum and maximum amplitude, zero-crossing and autocorrelation measurements. With these detectors, it is assumed the formant structure has been minimized by preprocessing the speech. A second category for pitch detection algorithms uses frequency-domain properties of speech signals. A periodic signal in the time-domain will consist of a series of impulses in the frequency-domain located at the fundamental frequency and its harmonics. Therefore, one can make measurements in the frequency domain to determine the pitch period. The final group combines both time and frequency-domain concepts of the speech signals in order to determine pitch period. This is a technique that is used which flattens the signal with frequency-domain techniques and subsequently uses autocorrelation measures to estimate the pitch period. These are called hybrid techniques. Previous work of the pitch detection algorithms and related works that have been published will be discussed.

There are several documented pitch extraction methods that have been published recently. In earlier methods, analysis of the speech time waveform were attempted by visual inspection of spectrograms which involved the manual determination of pitch [41]. At this time the authors noted the requirement for an automatic scheme of some kind. Pinson [42] used the method of Mathews, Miller, and David [41] to estimate a time-domain

synchronous pitch which in turn was used to determine frequencies and bandwidths of vowel formants.

Sondhi [43] introduced three methods for finding the pitch period. The first method spectrum flattens the signal and corrects the phase to synchronize harmonics. A second method by Sondhi also flattens the spectrum but adds an autocorrelation to determine pitch. The third method center clips the speech signal and uses autocorrelation for determination of pitch. Using the method by Sondhi, a real-time digital hardware pitch detector was implemented by Dubnowski, Schafer, and Rabiner [44].

There are also methods that make use of the power spectrum in the determination of the pitch. One such method is called cepstrum pitch determination. The cepstrum is defined as the power spectrum of the logarithm of the power spectrum, or mathematically expressed, the cepstrum, $Q(\tau)$ [45] [46], is

$$Q(\tau) = [\int_0^\infty \log|F(\omega)|^2 \cos(\omega\tau)\, d\omega]^2 \qquad (1.1)$$

where $f(t)$ is the speech signal, ω is the frequency in radians, and

$$F(\omega) = \int_{-\infty}^{\infty} f(t)\, e^{-j\omega t}\, dt \qquad (1.2)$$

More recently, using digital inverse filtering techniques, Markel has innovated a method for estimating the fundamental frequency of voiced speech using time-domain analysis. This method has been referred to as a simplified inverse filter tracking (SIFT) algorithm [47]. The pitch period is estimated by an interpolation of the autocorrelation function in the neighborhood of the peak of the autocorrelation function.

Another recent algorithm that determines the fundamental frequency of sampled speech is implemented by segmenting the signal into pitch periods. This is done by identifying the beginning of each pitch period. This algorithm is called the data reduction pitch detector by Miller [48]. To obtain the appropriate identity of the beginning of the pitch period, the method detects the cycles of the waveform based on intervals between major zero crossings. The rest of the algorithm determines principal cycles, which correspond to true pitch periods.

In work presented by Gold [49], it is assumed that pitch extraction could be obtained by a visual inspection of the speech wave and is the best obtainable. The computer program contains essentially four sections. First, a voiced/unvoiced decision is made and the two portions are separated. Each voiced portion is labeled as relative maximum, then the peak detector is compiled. The third decision is to determine the spacing; this in turn determines which samples will be called pitch peaks. Finally, a procedure is necessary to eliminate spurious peaks and add into the speech missing pitch peaks. The program is implemented such that editing can make the best pitch selection.

The work of Gold and Rabiner [50] using parallel processing for estimating pitch is a modified version of Gold [49]. A series of measurements are made to find the peaks and valleys of the signals. There are six cases used to determine this. Each is followed to determine if the sample will be an impulse or zero. The rules of this are:

1. An impulse equal to the peak of the signal occurs at the point of each peak in time.

2. An impulse equal to the difference between the signal present peak and the past peak amplitude occurs at the point of each peak in time.

3. An impulse equal to the difference between the signal present peak and the past peak amplitude occurs at the point of each peak in time. (If the difference is negative, then it is set to zero.)

4. An impulse equal to the negative of the peak of the signal occurs at each negative peak in time.

5. An impulse equal to the negative of the peak at each negative peak plus the peak of the preceding negative peak occurs at each negative peak in time.

6. An impulse equal to the negative of the peak at each negative peak, plus the negative of the preceding local minimum occurs at each negative peak. (If this difference is negative, then the impulse is set to zero.)

From this technique six estimates are formed. These estimates are combined with the two most recent estimates for each of the six pitch detectors. The values are then compared within an acceptable tolerance; the decision is made for the most occurrences. This value is declared the pitch at that time. An unvoiced decision is made when there is an inconsistency between the comparisons for the pitch period.

Another method by Atal [51] is based upon LPC. This detector initializes with a voiced/unvoiced decision. Upon being classified as voiced, the speech is low-pass filtered and then decimated by five to one. The method uses a 41-pole LPC analysis on 40 ms seconds of frame data to generate the speech harmonics. Then, a Newton transformation is used to spectrally flatten the speech. A peak picker determines the pitch period at the five to one decimated rating. Then, the signal is interpolated and a higher resolution is used to obtain the pitch period.

The average magnitude difference function (AMDF) pitch extractor [52] is a variation of autocorrelation analysis to determine the pitch period of voiced speech sounds. This method takes advantage of the periodicity of voiced speech. It calculates a difference function that at multiples of the pitch period will dip sharply when the delayed speech and original speech are compared. The AMDF function is implemented with subtraction, addition, and absolute value operations, whereas autocorrelation methods use addition and multiplication operations. For this reason, the AMDF function is attractive for real-time operations.

Another real-time pitch extraction method, based on linear predictive techniques, is presented by Maksym [53]. The method employs a nonstationary error process from the adaptive predictive coder by Atal [5]. The algorithm in addition to pitch period extraction also detects voiced speech. The basis of the method uses a predictive one-bit quantizer with an adaptive algorithm for determining prediction coefficients. Since the method operates on the short-term prediction of the speech waveform, the presence of the glottal excitation can be detected.

A semiautomatic pitch detector (SAPD) [54] has been presented by McGonegal, Rabiner, and Rosenberg. This method semiautomatically determines the pitch contour of an utterance. An autocorrelation of the speech is generated. The cepstrum of the unfiltered speech is computed. These displays are shown on a scope on a frame-by-frame basis. The computed pitch period for each waveform is marked by and is displayed to the user. With the incorporation of the three waveforms, an extremely accurate measure is found. The processing is lengthy for an utterance; however, robustness and accuracy of the results can be a trade-off for many applications.

A recent method for estimating pitch period in the presence of noise of voiced sounds is based on a maximum likelihood formulation [55]. This scheme is designed to be resistant to white, Gaussian noise. A new signal is formed from the speech signal with a maximizing function to enhance the peaks for short periods. The function is formed by an autocorrelation of the speech. It provides accurate estimates of the pitch period and can be used to determine formant structure. It is compared with the cepstrum method to perform better under the white noise conditions.

An automatic pitch extraction method was developed by Markel [56] which also determines formant frequency tracking. This method is similar to the cepstral analysis. The technique uses two FFT's to obtain the sequence from which the pitch is extracted. The difference between this method and the cepstral method is the procedure for determining the voiced/unvoiced decision.

An accurate method based on the prediction residual is the method by Atal and Hanauer [5]. The speech is low-pass filtered and each sample is raised to a third power to emphasize the high amplitudes of the speech waveform. A pitch-synchronous correlation analysis is performed of the cubed speech. A voiced/unvoiced decision is made in this technique. A second method is based on a linear prediction representation of the speech waveform. Each sample is predicted from the previous n samples, and therefore the correlation is not good at the beginning of the pitch period. The error is large at the beginning. The basis of the technique is to use peak picking for the pitch detection.

Another accurate method has been described by Itakura and Saito [57]. This method determines the prediciton error signal by the method of lattice filter formulation. The pitch period is determined by computing

autocorrelation coefficients of the residual. A set threshold compares the autocorrelation for a voiced/unvoiced decision with the pitch period.

A two stage method was developed by Boll [58] to determine the pitch period. The method is based on the Itakura [57] algorithm. It is built by adding the initialization of each frame based on the preceeding frame results. The portion of the autocorrelation function of the residual in the range where a pitch pulse is expected and the basis of the a priori information is computed in each frame. The savings in computation is significant.

Two methods were developed by Barnwell and others [59]. These algorithms are: 1) the multiband pitch period (MBPP) estimator, and 2) the skip-sample recursive least squares pitch position estimator. The multiband pitch period estimator first filters the speech waveform into four bands across the frequency regions where a fundamental is expected to occur. The bandwidths of these filters are chosen so that only one of the outputs will be expected to contain the fundamental. Zero-crossing pitch detectors operate on the outputs of each of the filters. The information derived from the zero-crossing detectors is used as a basis for logical operations to produce pitch period estimates. The skip sample recursive least squares technique is based on a recursive least squares linear predictive coder. The coder operates on a lower sampling rate than a linear predictive coder and it uses fewer coefficients than the predictive filter. This approach permits the original sampling time resolution to be retained. The method produces a sharp residual signal whose pitch pulses can be used to determine the period.

The future trend is towards efficient low-bit rate coding that enhances the perceptual quality and intelligibility of speech. The coding

of the residual signal is one way of arriving at the desired goal. This thesis presents such an idea along with a novel approach to pitch extraction. The next section presents the organization of the thesis.

1.3 Organization of the Thesis

Chapter II presents the basic ideas associated with the concept of the prediction residual. A discussion of the mechanism of speech production as related to the makeup of speech articulation is presented in speech science terms. A model of the vocal tract is presented in mathematical terms and the residual is presented in an algorithm form. The method of short-time analysis is presented. A new method for determining pitch implementation is presented using the residual waveform as the source function.

Chapter III presents some of the general ideas associated with coding of speech along with some applications. The method of transform coding (TC) is compared to the method of sub-band coding (SBC). The equivalence of the two methods is shown under certain conditions. The Articulation Index (AI) and the phoneme transitional information related to speech intelligibility are discussed along with their incorporation into the coding scheme to enhance the perception of speech. The results of the distribution of energy from the prediction residual of the phonemes are presented.

Chapter IV presents the design of the energy based sub-band coding algorithm. The basic ideas associated with the sub-band coding are discussed as related to the proposed coding scheme. The adaptive quantization is presented to explain the allocation of bits. The result on

signal-to-noise ratio (SNR) performance measurements are presented. The computation for coding the prediction residual is presented.

Chapter V presents a summary and suggestions for further study. The appendixes give a sample of the related speech science definitions, computer programs for coding the prediction residual, a brief review of the concept of Articulation Index and sonagrams of speech data.

CHAPTER II

PREDICTION RESIDUAL AND THE PITCH EXTRACTION

2.1 Introduction

Recent work in the area of speech analysis and synthesis is based upon a model that separates the glottal flow from the vocal tract. That is, the speech production is represented by a convolution model where the input corresponds to the glottal volume velocity and the vocal tract by a filter. Recent models have assumed an all pole filter to represent the vocal tract [5]. The filter coefficients are determined by using the method of linear prediction. By using the inverse filter, the speech can be deconvolved to obtain the prediction error or residual. The block diagram representing this is shown in Figure 3. The residual produces a peak where the prediction is bad, representing pitch period designations. As the prediction becomes more accurate, the residual appears as a noisy signal.

Most synthesis models use a filter excited by either a train of quasi-periodic pulses or a random noise source [60]. The periodic source excites the filter for voiced sounds. The noise source excites the filter for unvoiced sounds. The prediction residual is applicable for voiced or unvoiced sounds because the residual is an approximate signal of the corresponding input sources that generate these sounds. The detailed description of the prediction residual is discussed in Section 4 of this chapter.

Figure 3. Prediction Residual Formed by Speech Through an Inverse Filter

The linear predictive techniques described so far have been used successfully for time-domain speech analysis and synthesis [5] [30]. The linear predictive coding (LPC) techniques have been used in communications in the past; however, it was applied to speech only recently [5] [7]. The use of linear prediction in describing the transfer function of the vocal tract avoids the complexity of Fourier analysis. The slowly time varying aspects of speech can be taken into consideration by updating the filter coefficients every so often.

Two significant contributions have been made by Weiner [61] [62] and Shannon [63]. Weiner's work describes prediction and filtering of random, time series data. Shannon's results describe the information content of a message, related to band-width and time requirements of that message, related to band-width and time requirements of that message. The background of this chapter uses Weiner's method as applied to stationary data. Shannon's results are implicitly used in the coding scheme.

Section 2.2 describes the basis of human speech production. Section 2.3 discusses the vocal tract model as a discrete time invariant linear filter. Section 2.4 describes a parallel between the glottal waveform and the residual signal. Section 2.5 reviews linear prediction analysis. Section 2.6 discusses short-time analysis. Section 2.7 describes the implementation of operations for the calculation of the prediction residual. Section 2.8 presents a novel pitch extraction technique.

2.2 Mechanism of Speech Production

Man's system of communication is by speech. Speech is produced through the human vocal system in a continuous fashion. However, speech

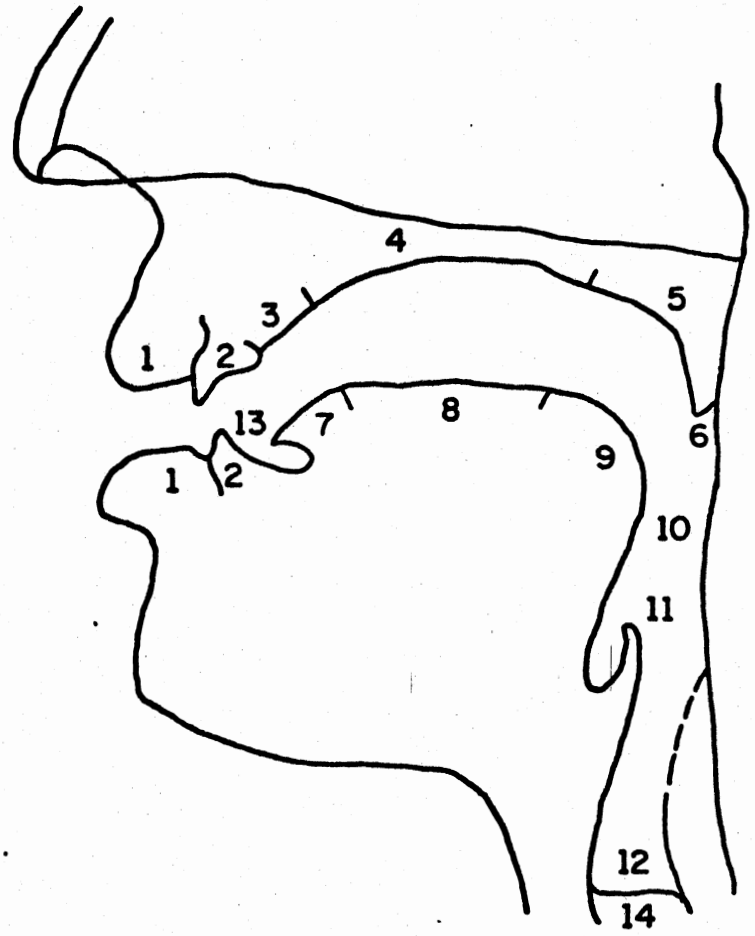
signals are composed of a sequence of discrete sounds called phonemes. Although phonemes are not bricks, they are the basic sounds that serve to make a complete word in any language. The connection or arrangement of these sounds is based on certain rules. It is the study of these rules and the way these sounds fit together that is called linguistics. The basic linguistic element is called a phoneme. Its distinguishable variations are called allophones [2].

Speech in humans is produced by a physical acoustic system consisting of principally four parts: lungs, vocal tract, nasal tract and vocal cords (see Figure 4). The lungs supply the volume of air necessary to produce speech. The vocal tract and nasal tract act as filters to shape the waveform. The velum, a small flap of skin, acts as a switch to close the entrance to the nasal tract. When closed, it removes any effect the nasal tract may have on the sound produced. The vocal cords, tongue, teeth and palate are parts of the filter or constriction mechanism. An elongated opening between the folds of the skin which make up the vocal cords is called the glottis.

The vocal tract provides the column of air, which is set to vibration by the excitation of the glottis. In an average male, the vocal tract is about 17 centimenters in length. The cross-sectional area which is determined by the position of the tongue, lips, jaw and velum varies from zero, i.e., complete closure, to approximately 20 square centimeters.

Speech sounds produced by the system can be separated into three distinct classes according to their mode of excitation. The voiced sounds are produced when air is permitted to escape in quasi-periodic pulses by the vibratory actions of the vocal cords. This sets the acoustic system to vibrating at its natural frequencies. These resonant

VOCAL SYSTEM (CROSS SECTIONAL VIEW)



1 - LIPS
2 - TEETH
3 - TEETH RIDGE
4 - HARD PALATE
5 - SOFT PALATE (VELUM)
6 - UVULA
7 - BLADE OF TONGUE
8 - FRONT OF TONGUE
9 - BACK OF TONGUE
10 - PHARYNX
11 - EPIGLOTTIS
12 - POSITIONS OF VOCAL CORDS
13 - TIP OF TONGUE
14 - GLOTTIS

Figure 4. Cross-Sectional View of the Human Tract System

frequencies are concentrations of energy and are known as formant frequencies. These are useful in characterizing the vocal tract configuration, as there is a one-to-one correspondence in the relationship of vocal tract configuration and formant frequencies. The fricative or unvoiced sounds are generated by forming a constriction at some point along the vocal tract and forcing air through the constriction at a velocity high enough to produce turbulence. This can be identified as wideband noise exciting the vocal tract. For an unvoiced sound the vocal cords are relaxed and partially open. The plosive sounds result from a complete closure of the vocal tract and a sudden or abrupt release of the closure.

The formants or natural resonances are numbered F_1, F_2, F_3, ... Typically, for speech analysis, only the first three or four are used. Table I gives representative values of these for certain vowels. It has been noted that all phonemes characterize some formant structure; however, it is most noted for voiced sounds [2]. It is indicative of the first formant to be greater in frequency than the fundamental frequency of the vocal tract. The fundamental frequency is the rate of vibration of the vocal cords; whereas, the first formant represents the first concentration of energy of the vocal tract system excited at the fundamental frequency. Typically, the fundamental frequency is around 120 Hertz for men, 220 Hertz for women and 300 Hertz for children. The pitch period is the reciprocal of fundamental frequency. The pitch period has a range from three milliseconds to eight milliseconds for voiced sounds. For the unvoiced sounds, most frequencies range above 4000 Hz and it has approximately a flat spectrum. All voiced sounds are characterized by voice onset time (VOT). For example, plosives are characterized by VOT, which is

TABLE I

AVERAGES OF FUNDAMENTAL AND FORMANT FREQUENCIES
AND FORMANT AMPLITUDES OF VOWELS BY 76 SPEAKERS

		i	I	ɛ	æ	ɑ	ɔ	ʊ	u	ʌ	ɝ
Fundamental frequencies (cps)	M	136	135	130	127	124	129	137	141	130	133
	W	235	232	223	210	212	216	232	231	221	218
	Ch	272	269	260	251	256	263	276	274	261	261
Formant frequencies (cps)											
F_1	M	270	390	530	660	730	570	440	300	640	490
	W	310	430	610	860	850	590	470	370	750	500
	Ch	370	530	600	1010	1030	680	560	430	850	560
F_2	M	2290	1990	1840	1720	1090	840	1020	870	1190	1350
	W	2790	2480	2330	2050	1220	920	1160	950	1400	1640
	Ch	3200	2730	2610	2320	1370	1060	1410	1170	1590	1820
F_3	M	3010	2550	2480	2410	2440	2410	2240	2240	2390	1690
	W	3310	3070	2990	2850	2810	2710	2680	2670	2780	1960
	Ch	3730	3600	3570	3320	3170	3180	3310	3260	3360	2160
Formant amplitudes (db)	L_1	-4	-3	-2	-1	-1	0	-1	-3	-1	-5
	L_2	-24	-23	-17	-12	-5	-7	-12	-19	-10	-15
	L_3	-28	-27	-24	-22	-28	-34	-34	-43	-27	-20

Source: Peterson and Barney, "Control Methods Used in a Study of the Vowels," The Journal of the Acoustical Society of America, Vol. 24, No. 2 (1952), 181.

the delay from complete closure of the plosive to the beginning of voicing [66]. The VOT ranges from 25 milliseconds to 300 milliseconds depending on the phoneme.

Each phoneme has its own characterization depending on the language. This characterization is associated with place of articulation and voicing. In this thesis, discussed are the phonemes of the English language. This is not to discard the pitch inflections in Chinese, whispered vowels in Japanese or vocal clicks of South African Hottentots, but to restrict to a basic area to all languages. This is established by the International Phonetic Association (IPA). Most linguists use about 35 basic units, and six diphthongs or combination phonemes. The symbols and teletype representations of these are shown in Table II.

Phoneticians classify speech sounds by vowels and consonants, or strictly speaking in the manner and their place of production. Each phoneme has certain characteristics and is identified from the distinctive features of the speech sound. The distinctive features give a unique identification of the phoneme. These are given below [68].

1. Vocalic/Nonvocalic

   presence vs. absence of a sharply defined formant structure.

2. Consonant/Nonconsonant

   low vs. high total energy.

3. Interrupted/Continuant

   silence followed and/or preceded by spread of energy over a wide frequency region (either as a burst or a rapid transition of vowel formants) vs. absence of abrupt transition between sound and the silence.

4. Nasal/Oral

TABLE II

REPRESENTATION OF IPA PHONEMES WITH EXAMPLES

Standard IPA	Teletype Representation	Example
i	IY	beet
I	IH	bit
e	EY	gate
ɛ	EH	get
æ	AE	fat
a	AA	father
ɔ	AO	lawn
o	OW	lone
∪	UH	full
u	UW	fool
ɝ, ɤ	ER	murder
ɑ	AX	about
ʌ	AH	but
aI	AY	hide
aU	AW	how
ɔI	OY	toy
p	P	pack
b	B	back
t	T	time
d	D	dime
k	K	coat

TABLE II (Continued)

Standard IPA	Teletype Representation	Example
g	G	goat
f	F	fault
v	V	vault
θ	TH	ether
ð	DH	either
s	S	sue
z	Z	zoo
ʃ	SH	leash
ʒ	ZH	leisure
h	HH	how
m	M	sum
n	N	sun
ŋ	NX	sung
l	L	laugh
w	W	wear
j	Y	young
r	R	rate
tʃ	CH	chan
d	JH	jar
hw	WH	where

Source: Rabiner and Schafer, Digital Processing of Speech Signals, New Jersey: Prentice-Hall, 1978, p. 43.

spreading the available energy over wider vs. narrower frequency regions by a reduction in the intensity of certain (primarily the first) formants and introduction of additional (nasal) formants.

5. Tense/Lax

   higher vs. lower total energy in conjunction with a greater vs. smaller spread of the energy in the spectrum and in time.

6. Compact/Diffuse

   higher vs. lower concentration of energy in a relatively narrow, central region of the spectrum accompanied by an increase vs. a decrease of the total energy.

7. Grave/Acute

   concentration of energy in the lower vs. upper frequencies of the spectrum.

8. Flat/Plain

   flat phonemes in contra-distinction to the corresponding plain ones are characterized by a downward shift or weakening of some of their upper frequency components.

9. Strident/Mellow

   higher intensity noise vs. lower intensity noise.

A table for the distinctive features of the phonemes of English are shown in Figure 5 [66]. As indicated above the features may be of two types. The presence or absence of each feature is expressed as a plus (+) or minus (-). For example, the vocalic category has vowels shown as plus and consonants are shown as minus.

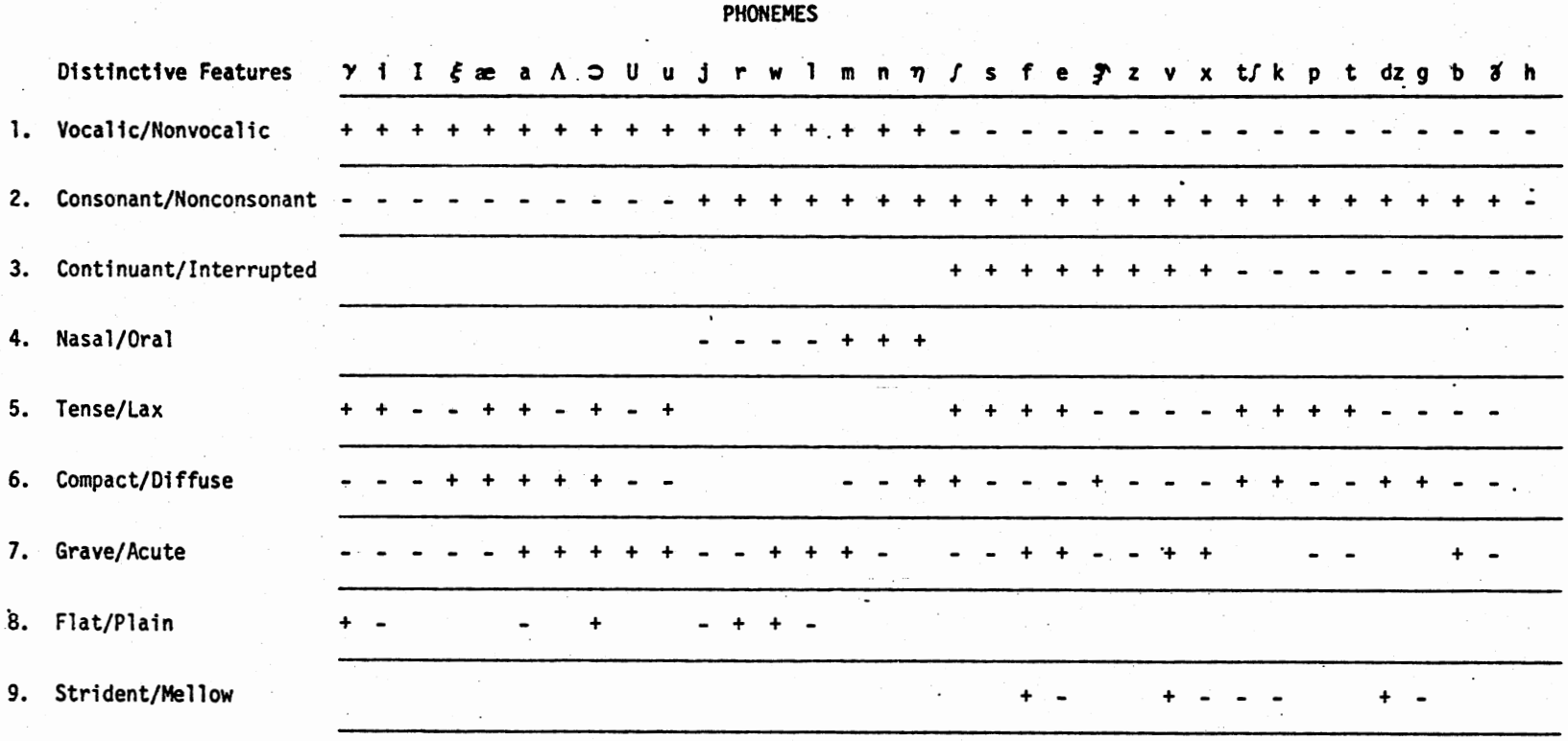


Figure 5. Distinctive Features of the Phonemes of English Indicating the Presence or Absence of a Feature

2.3 Model of the Vocal Tract

The acoustic speech system was qualitatively described in the previous section. The acoustic tube model of the vocal tract filter can be represented as a discrete time-invariant linear filter. The modeling has been discussed in the literature [2] [7] [67]. The acoustic tube is approximated by a number of sections each having a constant cross-sectional area. The cross-sectional area is characterized by the reflection coefficients. The reflection coefficient is the percentage of a wave reflected at an acoustic tube junction. The number of sections in the acoustic tube model is related to the number of formants for a phoneme.

The formants of speech correspond to the poles of the vocal tract transfer function [67]. As pointed out in the last section, only the first three or four formants are used for speech analysis, and these frequencies are below 5000 Hz. Generally, vocal tract resonances occur about one per thousand Hertz [67]. Therefore, a bandwidth of 5 kHz is, in general, sufficient for speech analysis and synthesis. Each phoneme is set apart from the others by the frequency location of the formants.

The majority of phonemes can be represented by an all-pole model of the vocal tract [5]. It is well known that for nonnasal voiced phonemes the transfer function of the vocal tract has no zeros [69]. Nasal and glide sounds include zeros in the transfer function. Zeros and poles are necessary to approximate the nasal and glide sounds. However, it has been shown that zeros in the vocal tract can be achieved by including more poles [5].

In Figure 1, let the transfer function of the vocal tract be expressed by [7]

$$V(z) = \frac{G}{1 - \sum_{i=1}^{P} a_i z^{-i}} \qquad (2.1)$$

where G, the gain; $\{a_i\}$, the filter coefficients, are a function of the cross-sectional areas of the acoustic tube. The value of P, the order of the system, is usually taken as twice the number of formants for analysis for each speech sound. Typical values for P range from 8 to 10. The value of 10 has been used for lattice network representations of the vocal tract.

It has been shown that given (2.1), a lossless tube model can be found [5] [7]. Also, given an acoustic tube with all areas positive, Equation (2.1) describes a stable system [7].

2.4 A Parallel Between Glottal Waveform and the Residual Signal

In modern signal processing techniques, it is necessary to use as much information as can be obtained about the structure of the signal. This section discusses the characteristics of the residual signal, which is the output of the linear prediction filter. It is the difference between the actual and predicted speech signals.

The residual signal used in this thesis is obtained by using the autocorrelation method in the LPC algorithm. In doing this, the speech is Hamming windowed, where the window function is

$$w(n) = 0.54 - 0.46 \cos\left[\frac{2\pi n}{N-1}\right] \qquad 0 \leq n \leq N-1$$

$$= 0 \qquad \text{otherwise} \qquad (2.2)$$

with N = 256. The computational details are discussed in Section 2.6.

The prediction residual is the ideal signal for the excitation function for LPC analysis [28]. It contains the actual information, rather than a pulse train or random noise as in the simplified linear prediction models [10]. The waveform that excites the vocal tract is the glottal waveform, and the residual approximates this.

The characteristics of the prediction residual are as follows: (1) it marks the pitch period, (2) it has basically a flat amplitude spectrum, (3) phasing information is embedded in the prediction residual, (4) the amplitude spectrum includes details related to the suprasegmentals of the individual and the spoken words, (5) the waveform includes the fact that voiced fricatives and stops are a combination of noise and a repetitive signal.

Figure 6 gives a comparison between a speech wave and the corresponding prediction residual for a particular phoneme. The computational aspects in obtaining these figures will be discussed later. The pitch period is marked by large spikes in the residual signal. The residual gives an excellent estimation of pitch since the glottal excitation is clearly marked.

Figure 7 displays an unsmoothed spectrum of the residual signal. The spectrum of the residual contains the formants also. The peaks of the formants are flattened; however, there is evidence of the fundamental and formant frequencies on the plot. The dashed line represents a smooth spectrum. Even in this, it is seen that there is evidence of the fundamental and the formants.

The pitch and voicing for each human is unique. It can be shown by spectrograms that individuals have unique voice prints. This uniqueness is basic to the excitation signal rather than in the vocal tract filter.

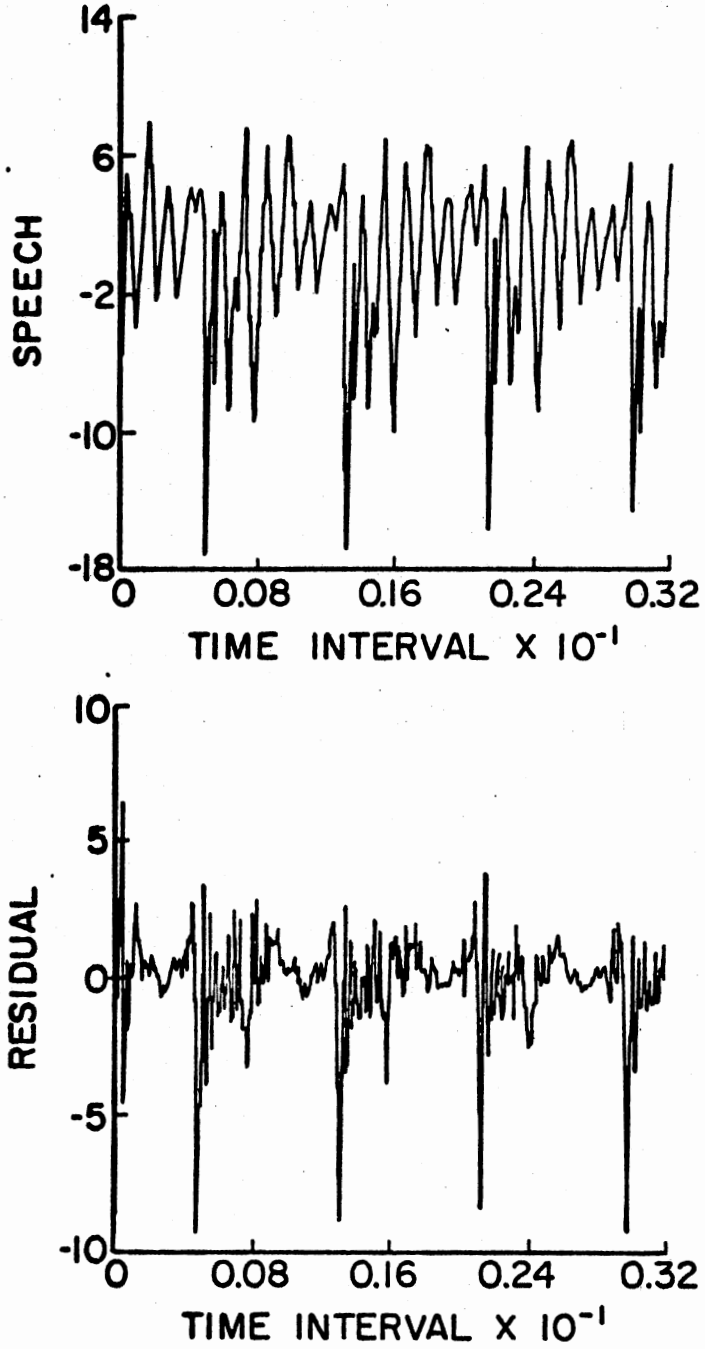


Figure 6. Speech Waveform and Its Prediction Residual for the Phoneme /æ/ over 256 Sample Interval

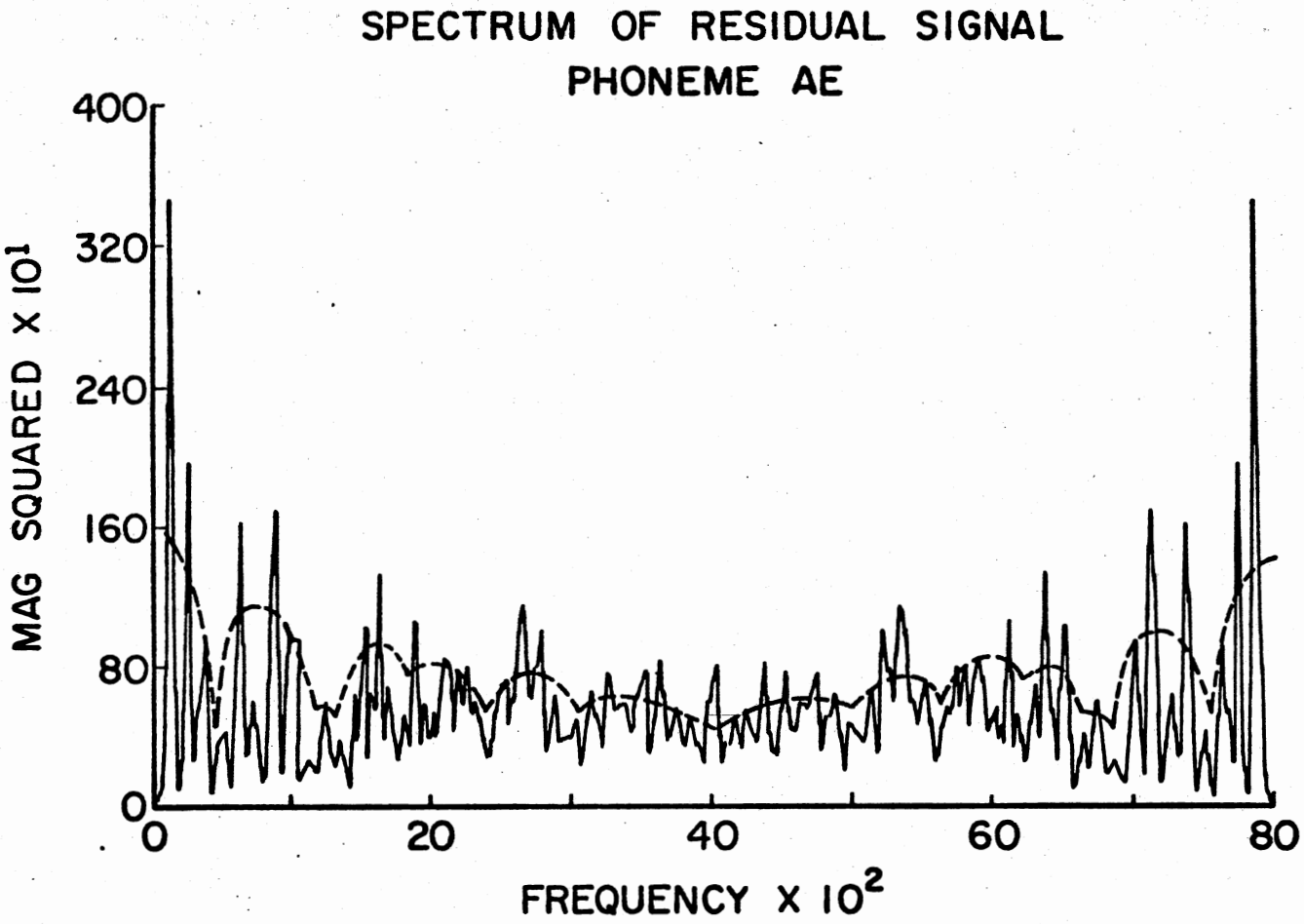


Figure 7. Spectrum of Residual Signal for the Phoneme /æ/

Therefore, the suprasegmentals, i.e., the intonation, dialect, melody pattern, etc., will remain unique to individuals for voiced sounds.

The voiced fricative lends more benefit to this discussion than its cognate, the unvoiced fricative. The unvoiced fricative is simply a noisy speech waveform that produces only a noisy residual signal. The fricative or stop is produced by forcing air through a constriction, such as the teeth or lips. The corresponding sound results from the turbulence and is of the noisy type. The waveform is then represented by noise that can be shown to be an unvoiced excitation source. However, the voiced fricative is a result of a constriction in the vocal tract while the vocal cords are vibrating. The residual signal from these phonemes produce a repetitive signal at the pitch period.

The artificial excitation function for voiced sounds result in speech that sounds a bit unnatural. The use of the prediction residual in coding methods would introduce naturalness in voicing. Ideally, the excitation of the vocal tract filter model should approximate the excitation of the human vocal tract. The prediction residual meets these requirements.

2.5 Review of Linear Prediction Analysis

Linear prediction analysis uses a weighted sum of P successive speech samples to predict the next speech sample. The weights are chosen such that the mean-square prediction error is minimized. Let

$$x_n \simeq a_1 x_{n-1} + a_2 x_{n-2} + \cdots a_P x_{n-P}$$

$$x_n \simeq \sum_{i=1}^{P} a_i x_{n-i} \qquad (2.3)$$

where x_n represents the speech sample sequence and a_i is a set of predictive coefficients. In this application, the method of least squares is used. Assuming a stationary linear system [5] with time-invariant statistics, zero mean, let $\hat{x}_f(n)$ represent the best estimate, in the least mean-square sense, of x_n using the a_i, $i = 1, \ldots, P$ coefficients and let $\hat{x}_b(n)$ be the best backward prediction of x_n using the b_i, $i = 1, \ldots, P$ coefficients. Then

$$\hat{x}_f(n) = \sum_{i=1}^{P} a_i \, x_{n-i} \tag{2.4a}$$

$$\hat{x}_b(n-P-1) = \sum_{i=1}^{P} b_i \, x_{n-i} \tag{2.4b}$$

Let $e_f(n)$ and $e_b(n)$ be the forward and backward prediction errors defined by

$$e_f(n) = x_n - \hat{x}_f(n)$$

$$= - \sum_{i=0}^{P} a_i \, x_{n-i} \tag{2.5a}$$

$$e_b(n-P-1) = x_{n-P-1} - \hat{x}_b(n-P-1)$$

$$= - \sum_{i=1}^{P+1} b_i \, x_{n-i} \tag{2.5b}$$

where it is assumed that $a_0 = -1$ and $b_{P+1} = -1$. Figure 8 gives the implementation of (2.5)

Since stationarity is assumed, it follows that the errors can be minimized by

$$E\left[\frac{\partial}{\partial a_j} \left(e_f(n)\right)^2\right] = 0 \qquad j = 1, \ldots, P \tag{2.6a}$$

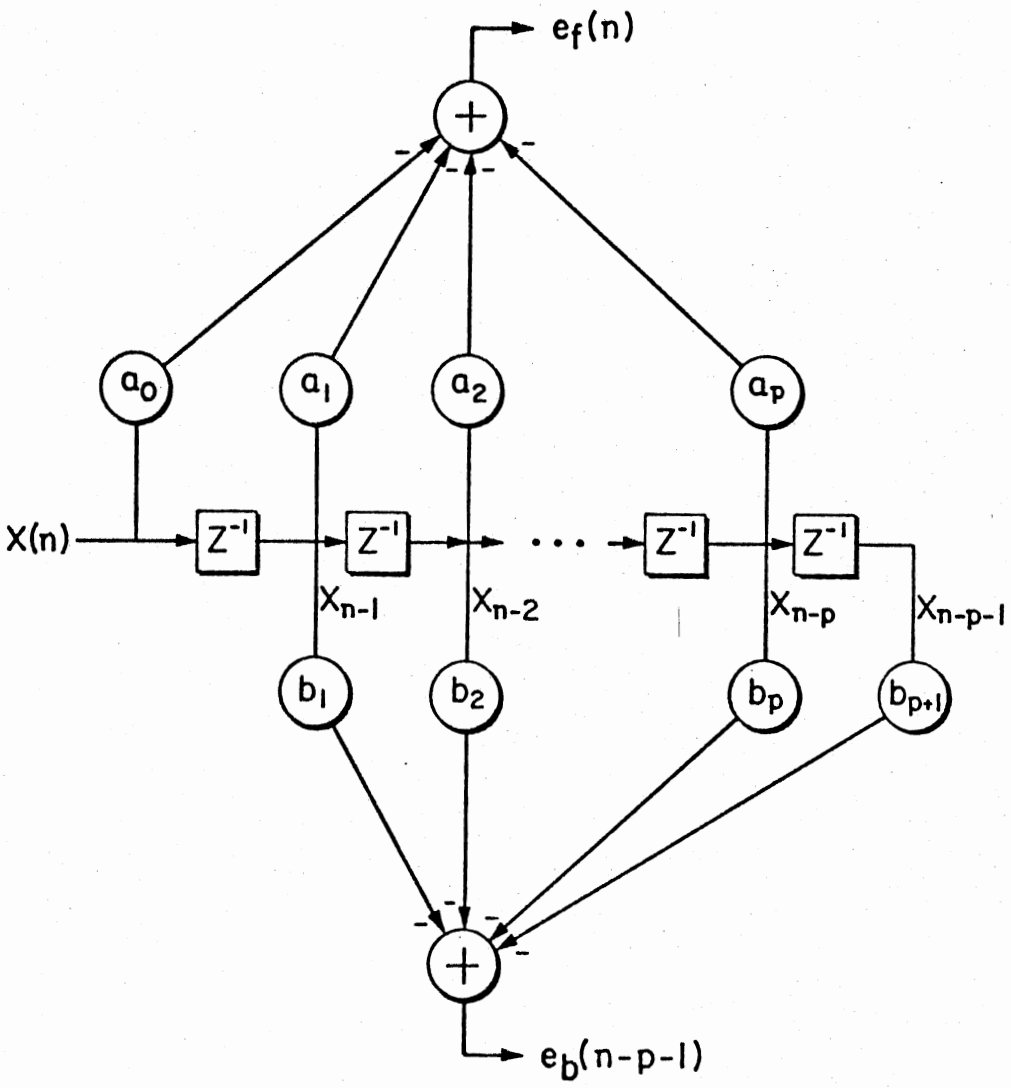


Figure 8. Implementation for Generation of Forward and Backward Prediction Errors

$$E\left[\frac{\partial}{\partial b_j}\left(e_b(n-P-1)\right)^2\right] = 0 \qquad j = 1, \ldots, P \tag{2.6b}$$

These reduce to

$$\sum_{i=0}^{P} a_i \, E[x_{n-i} \, x_{n-j}] = E[x_n \, x_{n-j}] \qquad j = 1, \ldots, P \tag{2.7}$$

$$\sum_{i=0}^{P} b_i \, E[x_{n-i} \, x_{n-j}] = E[x_{n-P-1} \, x_{n-j}] \qquad j = 1, \ldots, P \tag{2.8}$$

By defining

$$E[x_{n-i} \, x_{n-j}] = R_{i-j} \tag{2.9}$$

Equations (2.7) and (2.8) can be expressed by

$$\sum_{i=1}^{P} R_{i-j} \, a_i = R_j \qquad j = 1, \ldots, P \tag{2.10a}$$

$$\sum_{i=1}^{P} R_{i-j} \, b_i = R_{P+1-j} \qquad j = 1, \ldots, P \tag{2.10b}$$

where $R_{i-j} = R_{j-i}$ has been used. It is clear that (2.10a) and similarly (2.10b) can be written in a matrix form, wherein the coefficient matrix is a symmetric Toeplitz matrix [71]. Furthermore,

$$b_i = a_{P+1-i} \qquad i = 1, \ldots, P \tag{2.11}$$

which can be seen by defining

$$j = P+1-\ell$$

$$i = P+1-k$$

in (2.10a). That is,

$$\sum_{k=P}^{1} R_{k-\ell}\, a_{P+1-k} = R_{P+1-\ell}$$

which can be rewritten as

$$\sum_{i=1}^{P} R_{i-j}\, a_{P+1-i} = R_{P+1-j} \qquad j = 1, \ldots, P \tag{2.12}$$

Comparing (2.12) with (2.10), the relation in (2.11) can be seen The forward prediction error

$$E[e_f^2(n)] = E[(x_n - \sum_{i=1}^{P} a_i\, x_{n-i})(x_n - \sum_{j=1}^{P} a_i\, x_{n-i})]$$

$$= E[x_n^2 - 2 \sum_{i=1}^{P} a_i\, x_n x_{n-i}] + E[\sum_{j=1}^{P} a_j \sum_{i=1}^{P} a_i\, x_{n-i}\, x_{n-j}]$$

$$= E[x_n^2] - E[\sum_{i=1}^{P} a_i\, x_n\, x_{n-i}]$$

$$= R_o - \sum_{i=1}^{P} a_i\, R_i \equiv E_p \tag{2.13}$$

where (2.7a) has been used to obtain (2.13).

The cross-correlation between the forward and backward prediction errors is derived in the following. Let

$$C_{P+1} = E[e_f(n)\, e_b(n-P-1)]$$

$$= E[x(n)\, x(n-P-1)] - E[\sum_{j=1}^{P} a_{P+1-j}\, x_n x_{n-j}]$$

$$- E[\sum_{j=1}^{P} a_i\, x_{n-i}\, x_{n-P-1}] + E[\sum_{j=1}^{P} a_{P+1-j} \sum_{i=1}^{P} a_i\, x_{n-i}\, x_{n-j}]$$

$$= R_{P+1} - \sum_{i=1}^{P} a_i R_{P+1-i} \qquad (2.14)$$

where again (2.7a) has been used to obtain (2.14).

It is clear that (2.13) and (2.14) correspond to P coefficients. In the following, a recursive method will be used wherein the coefficients a_i will be updated. For this reason, let

$$E_o = R_o \qquad (2.15)$$

$$E_k = R_o - \sum_{i=1}^{k} a_i^{(k)} R_i \qquad k = 1, \ldots, P \qquad (2.16)$$

$$C_{k+1} = R_{k+1} - \sum_{i=1}^{k} a_i^{(k)} R_{k+1-i} \qquad k = 0, 1, \ldots, P-1 \qquad (2.17)$$

where $a_i^{(k)}$ are determined from (2.10a) by using

$$\sum_{i=1}^{k} R_{i-j} a_i^{(k)} = R_j \qquad j = 1, \ldots, k \qquad (2.18)$$

Durbin's method [72] [73] can now be used to solve for $a_i^{(k)}$ in (2.18). The corresponding equations are

$$E_o = R_o \qquad (2.19)$$

$$\left.\begin{aligned} k_{j+1} &= \frac{C_{j+1}}{E_j} \\ a_{j+1}^{(j+1)} &= k_{j+1} \\ a_i^{(j+1)} &= a_i^{(j)} - k_{j+1} a_{j+1-i}^{(j)} \\ E_{j+1} &= E_j (1 - k_{j+1}^2) \end{aligned}\right\} \quad \begin{aligned} &j = 0, 1, \ldots, P-1 \\ &i = 1, 2, \ldots, j \end{aligned} \qquad (2.20)$$

The predictive coefficients are obtained from

$$a_i = a_i^{(j+1)} \qquad i = 1, 2, \ldots, P$$

Interestingly, the prediction residual E_j in (2.20) is readily available in the algorithm for the predictor of order j. The coefficients k_j generated in (2.20) are usually referred as PARCOR coefficients. These have some interesting characteristics [9] [28].

1. $|k_j| \leq 1$
2. Since $|k_j|$ is unity bounded, a set quantization levels can be determined.
3. The PARCOR coefficients are the result of the orthogonalization of the auto-correcation matrix.

In order to show the application of this system, the transfer function and the algorithm to acquire the prediction residual is derived below.

The transforms of $e_f(n)$ and $e_b(n-P-1)$ in (2.5) can be expressed in terms of

$$E_f(z) = - \sum_{i=0}^{P} a_i z^{-i} X(z) \tag{2.21a}$$

$$z^{-(P+1)} E_b(z) = - \sum_{i=1}^{P+1} b_i z^{-i} X(z) \tag{2.21b}$$

where $E_f(z)$, $E_b(z)$ and $X(z)$ are the transforms of $e_f(n)$, $e_b(n)$ and $x(n)$, respectively. Note that a_o and b_{P+1} in (2.21) are each equal to -1. For simplicity, let

$$A_p(z) = 1 - \sum_{i=1}^{P} a_i z^{-i} \tag{2.22a}$$

$$B_p(z) = z^{-(P+1)} - \sum_{i=1}^{P} b_i z^{-i} \qquad (2.22b)$$

With these, (2.21) can be written as

$$E_f(z) = A_p(z) X(z) \qquad (2.23a)$$

$$E_b(z) = z^{P+1} B_p(z) X(z) \qquad (2.23b)$$

It is clear that (2.22) was implemented in Figure 8 using the direct form. Next, the lattice network implementation of (2.22) is discussed below. In order to do this, recall the relation $b_i = a_{P+1-i}$ given in (2.11). With the relation, (2.21b) can be written as

$$B_p(z) = z^{-(P+1)} - \sum_{i=1}^{P} a_{P+1-i} z^{-i}$$

$$= z^{-(P+1)} - \sum_{j=1}^{P} a_j z^{-(P+1)+j} \qquad (2.24)$$

$$= z^{-(P+1)} A_p(z^{-1}) \qquad (2.25)$$

From this it follows that

$$A_p(z) = z^{-(P+1)} B_p(z^{-1}) \qquad (2.26)$$

Equations (2.20), (2.25) and (2.26) will now be used to derive the lattice implementation. To develop the recursive equation for the lattice formulation, some of the above equations have to be written in a recursive manner. It is clear that (2.22) can be rewritten in the form

$$A_{j+1}(z) = - \sum_{i=0}^{j+1} a_i^{(j+1)} z^{-i} \qquad j = 0, 1, \ldots, P-1 \qquad (2.27a)$$

$$B_{j+1}(z) = -\sum_{i=1}^{(j+2)} b_i^{(j+1)} z^{-i} \qquad j = 0, 1, \ldots, P-1 \qquad (2.27b)$$

where the superscripts on a_i and b_i are included to denote that (j+1)th order is implemented rather than a Pth order. Also

$$a_o^{(j+1)} = -1 \qquad (2.28)$$

$$b_{j+2}^{(j+1)} = -1 \qquad (2.29)$$

have been used. The remaining $a_i^{(j+1)}$ can be expressed in terms of $a_i^{(j)}$ using (2.20); $b_i^{(j+1)}$ are related to $a_i^{(j+1)}$ by [see (2.4)]

$$b_i^{(j+1)} = a_{j+1-i}^{(j+1)} \qquad (2.30)$$

Using (2.20), (2.29) and (2.30) in (2.27a)

$$A_{j+1}(z) = 1 - \sum_{i=1}^{j+1} (a_i^{(j)} - k_{j+1}\, a_{j+1-i}^{(j)})\, z^{-i}$$

$$= A_j(z) - k_{j+1} \sum_{i=1}^{(j+1)} b_i^{(j)} z^{-i}$$

$$= A_j(z) - k_{j+1} B_j(z) \qquad (2.31)$$

Using (2.24)

$$B_{j+1}(z) = z^{-(j+2)} A_{j+1}(z^{-1}) \qquad (2.32)$$

Equation (2.31) can be rewritten as

$$A_{j+1}(z^{-1}) = A_j(z^{-1}) - k_{j+1} B_j(z^{-1}) \qquad (2.33)$$

Substituting (2.32) in (2.33) and simplifying

$$B_{j+1}(z) = z^{-1}[B_j(z) - k_{j+1} A_j(z)] \tag{2.34}$$

Equations (2.31) and (2.34) define the algorithm. The implementation of these is shown in Figure 9, where the generation of k_{j+1} is also included. The detailed structure of the optimum inverse filter as an analysis model is shown in Figure 10a. The corresponding synthesis model is shown in Figure 10b. The output of the synthesis filter is the input speech signal. From the analysis section, transform of the prediction residual is $A_p(z)$.

2.6 Short-Time Analysis

The concept of short-time Fourier analysis [76] [77] is fundamental for coding the residual signal. For a quasi-periodic signal such as speech, the short-time or time-dependent Fourier analysis allows for a detailed study.

The speech signal, x(m), m = 0, 1, ..., L-1, from Equation (2.3) is segmented into r sections such that short-time spectral analysis can be used. It is assumed that L = rN, where N corresponds to the number of samples in each section. This assumes the use of the formula

$$\sum_{n=-\infty}^{\infty} w(nD-m) = 1 \tag{2.35}$$

where w(m) corresponds to a band limited function to a frequency of 1/2D, and D is the period (in samples) between adjacent samples of the short-time transform of the signal [77]. In all practical cases, w(m) is a time limited signal and, therefore, its spectrum cannot be band limited. The effects of this non-band limited case are discussed in a recent paper [93]. It has been shown that the aliasing errors are small and can be

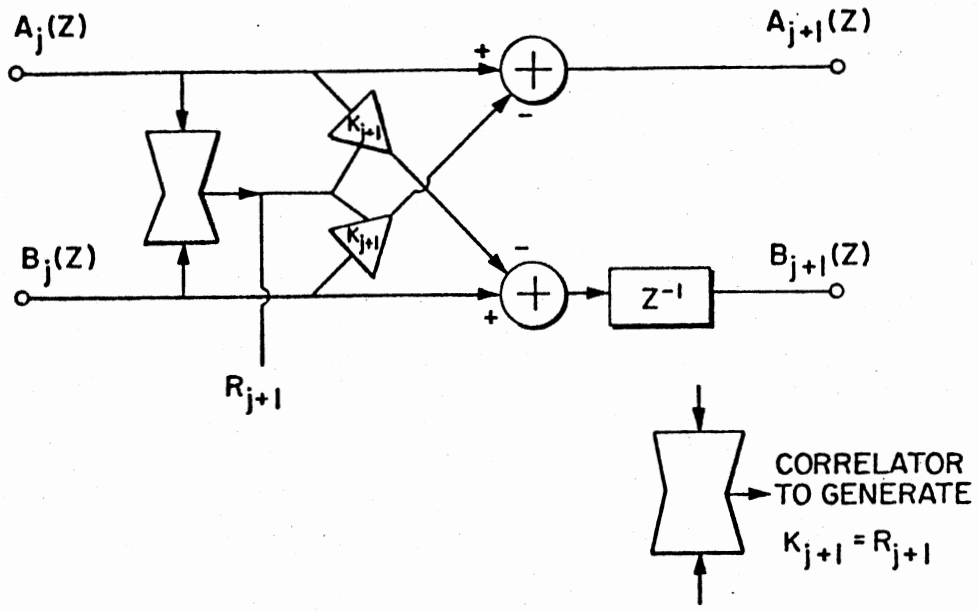
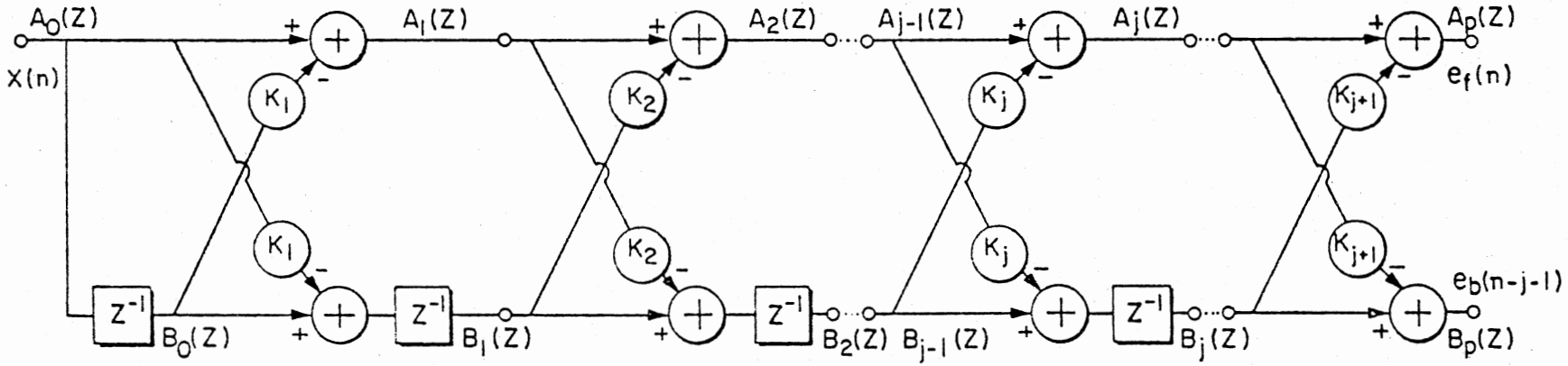
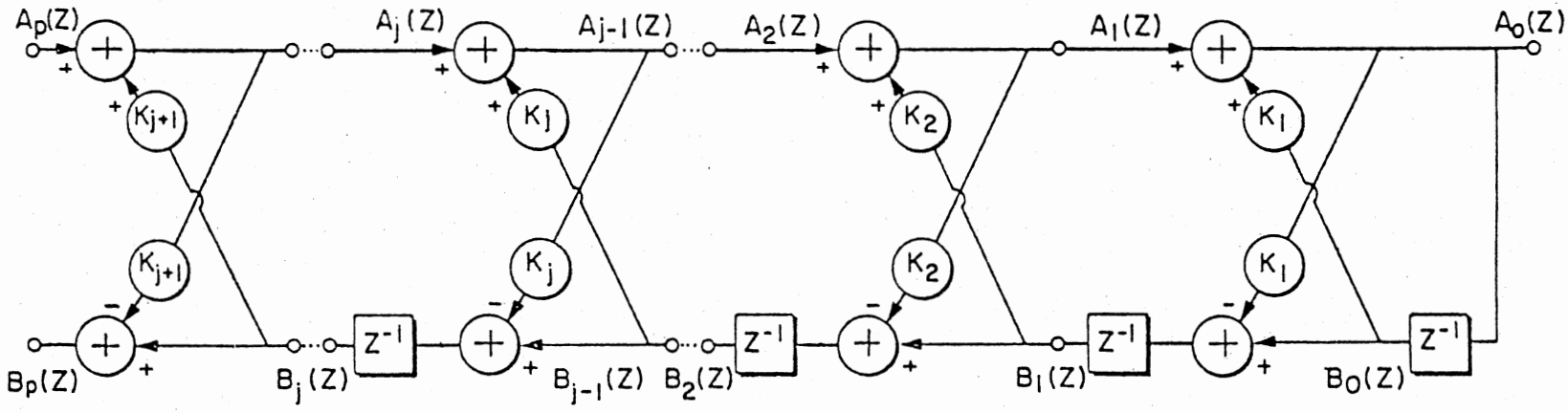


Figure 9. Detailed Structure of PARCOR Implementation for jth Stage



a.) Analysis Model



b.) Synthesis Model

Figure 10. Analysis and Synthesis Models for Lattice Structure

neglected if D is properly chosen. For a Hamming window, D = (N/4) [77]. In addition to the aliasing errors, end effect errors have to be considered also [94]. This is necessary since L if finite. The LPC analysis is applied to the windiwed signal resulting in the windowed residual signal. The overlap-add of this signal is the residual signal, which has error identified earlier.

The short-time Fourier transform of the residual signal $e_f(m)$ can be defined as [67]

$$X_n(e^{j\omega_k}) = \sum_{m=-\infty}^{\infty} [e_f(m)\, w(nD-m)]\, e^{-j\omega_k m} \qquad (2.36)$$

where $\omega_k = (2\,k/N)$, $k = 0, 1, \ldots, N-1$, and w(m) corresponds to a window. For a particular value of n, Equation (2.36) can be implemented using FFT. This is used in this thesis. A brief review of this is presented below.

Let

$$e_n(m) = e_f(nD+m)\, w(-m) \qquad -\infty \leq m \leq \infty \qquad (2.37)$$

Using this in (2.36),

$$X_n(e^{j\omega_k}) = \left[\sum_{m=-\infty}^{\infty} e_n(m)\, e^{-j\omega_k m}\right] e^{-j\omega_k n} \qquad (2.38)$$

Further, let $m = Nr+q$, $-\infty \leq r \leq \infty$, $0 \leq q \leq N-1$. With these,

$$X_n(e^{j\omega_k}) = \sum_{r=-\infty}^{\infty} \left[\sum_{q=0}^{N-1} e_n(Nr+q)\, e^{-j\omega_k (Nr+q)}\right] e^{-j\omega_k n} \qquad (2.39)$$

Noting that $e^{-j\omega_k Nr} = 1$,

$$X_n(e^{j\omega_k}) = \sum_{q=0}^{N-1} \left[\sum_{r=-\infty}^{\infty} e_n(Nr+q) \right] e^{-j\omega_k q} e^{-j\omega_k n} \qquad (2.40)$$

For simplicity, let

$$u_n(q) = \sum_{r=-\infty}^{\infty} e_n(Nr+q) \qquad 0 \leq q \leq N-1 \qquad (2.41)$$

Note that $u_n(q)$ is periodic with period N. Now Equation (2.40) can be written, and is

$$X_n(e^{j\omega_k}) = e^{-j\omega_k n} \left[\sum_{q=0}^{N-1} u_n(q)\, e^{-j\omega_k q} \right] \qquad (2.42)$$

Observe that $X_n(e^{j\omega_k})$ is represented as $e^{-j\omega_k n}$ times the DFT of the sequence $u_n(q)$. Therefore, (2.42) can be written as

$$e^{j\omega_k n} X_n(e^{j\omega_k}) = \sum_{q=0}^{N-1} u_n((m-nD))_N\, e^{-j\omega_k q} \qquad (2.43)$$

Equation (2.42) represents the DFT form, where $((\cdot))_N$ corresponds to the modulo N.

The following procedure can be used to compute (2.43).

1. The windowed sequence, $e_n(m)$, can be computed from (2.37). The sequence can then be divided into r sections of N samples each, where in this thesis, L = 4096, N = 256, D = 64, and r = 16.

2. The N-point DFT of $u_n((m-nD))_N$ can be computed to obtain (2.43) using FFT.

The above procedure is given here for generality. Due to the limitation of the disc space and to reduce computational time, a slightly different procedure is used in computing the spectral analysis. The residual signal is rectangular windowed to 256 points, spectrum analyzed

and then averaged. This is used only for phonemes discussed in the thesis. No overlapping was used. The errors associated with this method are quantified in previously mentioned references [93] [94].

2.7 Implementation of Operations for the Calculation of the Prediction Residual

In this section, the formulation of the prediction residual from the speech input is presented. The implementation of the operations to calculate the prediction residual represents the analysis model for LPC. The analysis model consists of the speech as the input, the vocal tract model, the correlation coefficients and the residual as the output.

The analog speech signal is band limited to 3600 Hertz using a second order Butterworth filter. This signal is digitized at the rate of 8000 samples/second. The algorithm for digitization is named DIGITIZ and the computer program is included in Appendix B.

The results in the last section are used to obtain the windowed digitized data. This allows to process the speech in short segments. The underlying assumption for most speech processing schemes is that the properties of the speech signal change relatively slowly with time [67]. This assumption leads to short-time methods which isolate the signal during the segment of windowing. The window is a 256-point Hamming window and is overlapped at 64-point intervals. The windowing is computed by program WINDOW in Appendix B.

The windowed signal is passed to program AUTO [7]. This program uses the autocorrelation method for solving the matrix equation (2.10) for the predictor coefficients [61]. The other matrix values solved for are the

reflection coefficients or PARCOR coefficients. These values are passed for use in the lattice formulation.

The lattice method represents a recursive algorithm for a solution of the prediction residual. This method guarantees stability. Note that the PARCOR coefficients are bounded. The program to calculate the residual by the lattice formulation is INVERS and is included in Appendix B.

Figure 11 illustrates a block diagram showing the sequence of operations related to the calculation of the prediction residual, $e_f(n)$.

2.8 A Novel Approach to Pitch Extraction

2.8.1 Types of Problems Associated with Pitch Extraction

The pitch extractor is of prime importance in most speech processing systems, as the pitch is one of the basic parameters in speech analysis and synthesis studies. In low-bit rate systems, it is an essential component [2] [7]. Speech with a constant fundamental frequency is perceived as a monotone or of a synthetic nature; variable pitch lends to speech a melody. An accurate pitch extractor is a challenging area of speech processing.

The difficulty in accurately determining pitch is due primarily to the time varying aspects of the glottal excitation. Since the model of the vocal tract assumes quasi-periodic changes occurring along the acoustic tube, the glottal response is not predicted accurately. This innaccuracy is due to the nonuniform train of periodic pulses that occur with the golttal waveform. The simple model of the vocal system excitation, i.e., periodic uniform pulses or Gaussian noise, eases the measurement of

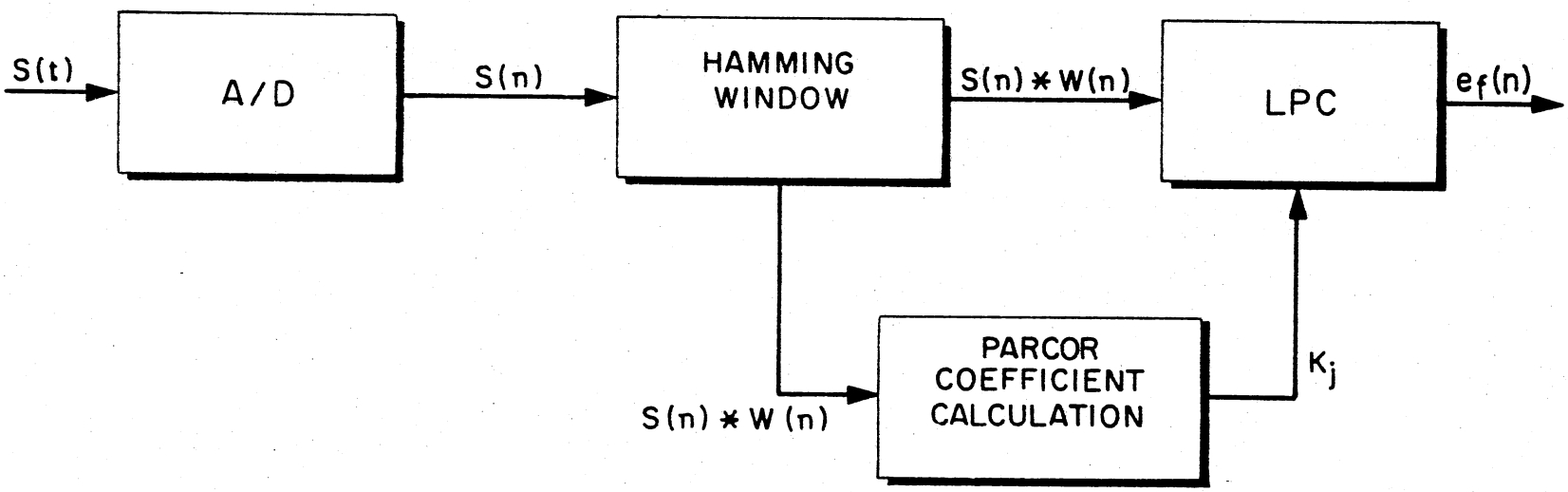


Figure 11. Sequence of Operations Related to Calculation of the Prediction Residual r(n)

of the period of the pitch. However, when the pitch and the waveform are changing within a period which occurs with frequency shifts, difficulty arises.

The second problem associated with the measurement of pitch is due to the nonseparability of the vocal tract model from the glottal excitation. That is, the separation of the formants and the fundamental frequency may not be possible and therefore the detection of the pitch period is difficult. This interaction can be seen most often during transitional regions of formants when the articulatory elements are changing.

The third problem is the detection of the beginning and ending of the pitch period. Part of this problem occurs in the definition of beginning and ending of the pitch period. In examining the speech waveform, it is necessary to always be consistent with the method because different definitions will often lead to different results. This is seen in Figure 12. In Figure 12, one can detect the period of zero crossings before the maximum peaks or detect the period between the maximums. However, the two methods do not always give the same answers. The discrepancy between the two is due to the slowly time-varying properties of glottal excitation.

The fourth problem that arises is the decision to ascertain which segment of speech is voiced or unvoiced. In particular, some algorithms have problems distinguishing between low-level speech and unvoiced speech. In transitional analysis, it is difficult to pinpoint the difference between the two.

In addition to the above problems, the pitch detection is hindered further when the signal is a transmitted speech signal. During the

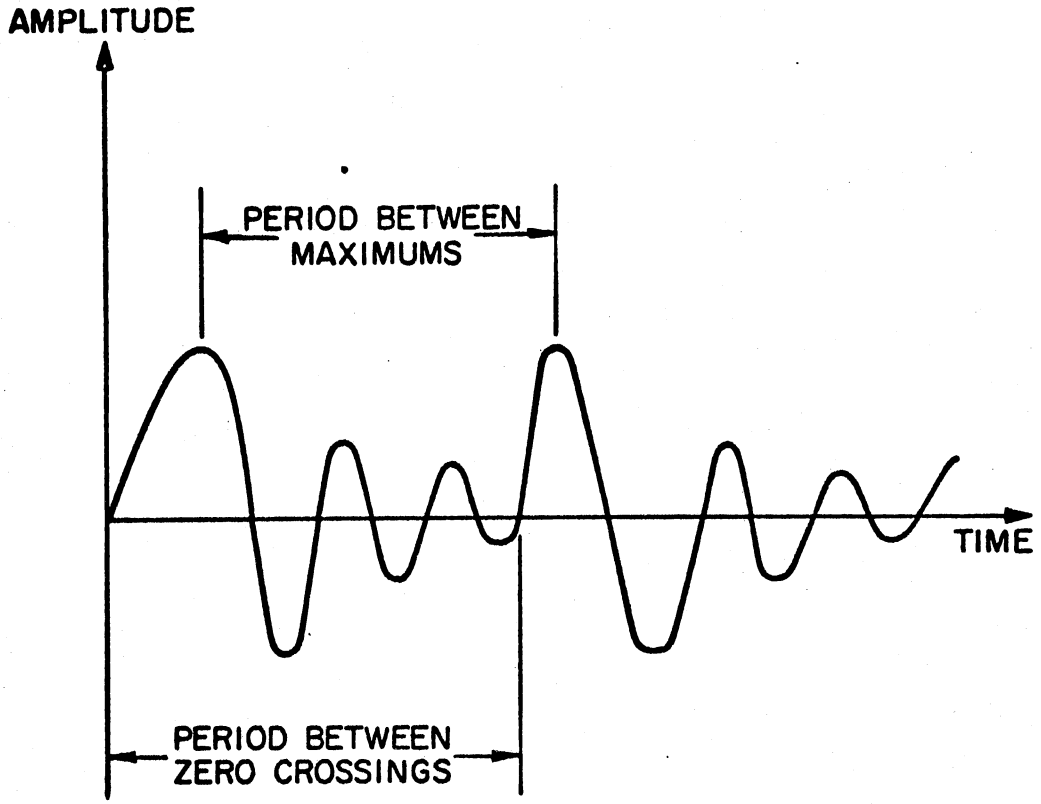


Figure 12. Two Methods of Determining Pitch Period

transmission of a speech signal over a telephone line, there are degradations that occur that can change the signal to make pitch detection difficult. These include: 1) phase distortion, 2) amplitude modulation of the signal, 3) crosstalk between messages, 4) clipping of high-level sounds. Furthermore, as the signal travels through the telephone lines, the lines act as a bandpass filter with approximate band edges f_1 = 200 Hertz and f_2 = 3200 Hertz. The fundamental frequency is usually less than 200 Hertz and therefore is removed by the bandpass action of the line. The pitch must be regenerated by using harmonics.

The next section discusses advantages and disadvantages associated with the use of the prediction residual for pitch extraction.

2.8.2 Advantages and Disadvantages for Using the Prediction Residual as a Source for Pitch Extraction

The prediction residual solves the problem of vocal tract excitation. Earlier, it is stated that there is inaccuracy in determining glottal response when using the simple model for excitation. When using the two-source model for the vocal system excitation, i.e., quasi-periodic pulses and random noise, a simple algorithm can be used for extraction of pitch. The residual can be used as a single source as an approximation to the glottal excitation, and, therefore, a simple method can be used to employ the residual to extract pitch.

It is well known that the residual represents the deconvolution of the speech from the vocal tract [7]. For each vocal tract configuration, a different set of formants and a variation in harmonics of the fundamental frequency in the spectrum is acquired. The pitch markings are

determined by residual spikes in the time-domain. This can be used to extract the pitch accurately.

The advantage of an accurate estimation of pitch will aid to the perception of speech. Any enhancement to perception is important to any analysis-synthesis speech system. The discussion which follows includes other reasons for using the prediction residual as a source for pitch extraction.

Referring to Figure 12, it is shown where errors can occur when the speech signal is used for pitch extraction. Figure 13 shows the residual, over 256 samples, characterized by spikes which represent the pitch period. It can be seen that it is not necessary to account for the zero crossings or maximums. It is simply a matter of tracking absolute maximums within the range of the established pitch period. It has been shown that if the interval of analysis is small enough the residual can be used to extract pitch accurately [28]. Future transmission rates will require a system that can do an acceptable performance for extracting pitch.

An application for using the residual signal for pitch extraction is with embedded coding. The advantage with the residual signal is that an absolute pitch can be determined in a frame. At higher transmission rates, the coding of the residual can be accomplished more efficiently. Therefore, a pitch extraction method can be employed easily. However, it is not feasible to transmit the residual with low transmission rates; consequently, the higher rates must extract the pitch and transfer this to the lower rates. Since the residual demonstrates a very accurate representation of the pitch, the frame-by-frame analysis of the pitch from the prediction residual would enhance pitch in an embedded coding scheme.

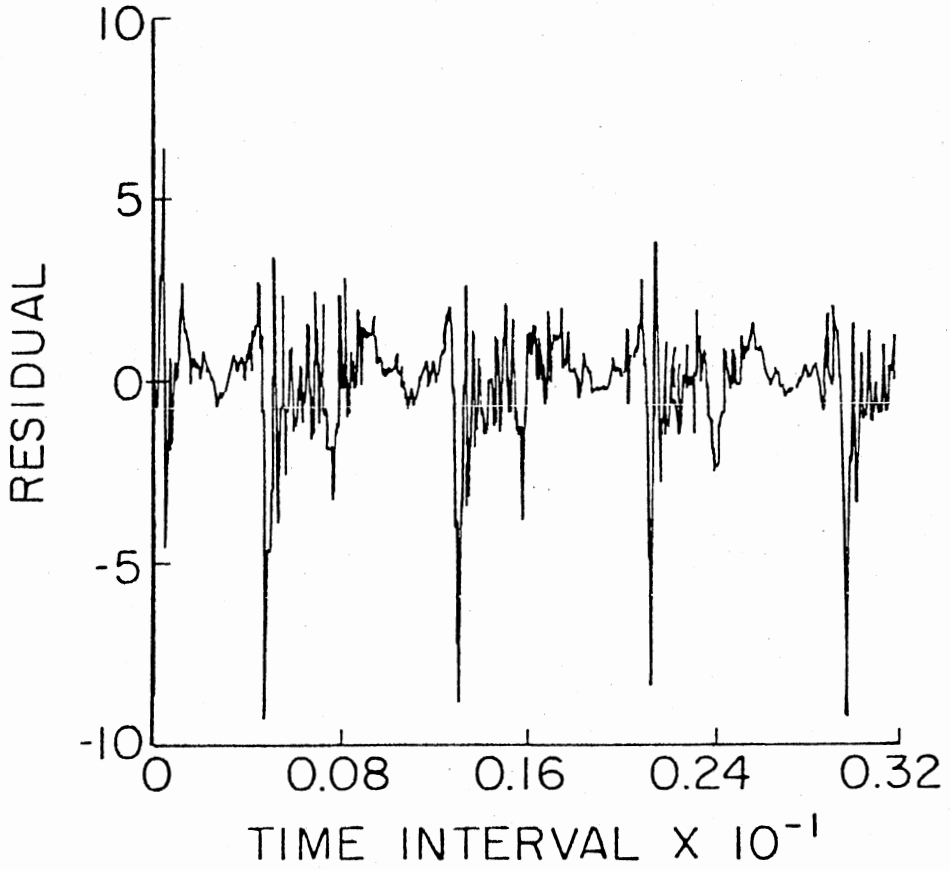


Figure 13. Prediction Residual Waveform for Phoneme /æ/ over 256 Sample Segment

However, a disadvantage associated with the residual may occur in a high noise environment. It has been shown earlier that the residual is a combination of periodic and noisy signals. In a high noise environment, the noise may overcome the residual signal. If the noise has amplitude in the range of pitch markings, the signal would require enhancement to extract pitch adequately. On the other hand, low noise contributes to the flatness of the spectrum of the signal and enhances pitch extraction.

Several advantages and disadvantages have been discussed. It can be readily seen that the residual is an ideal signal for extraction of the pitch. The next section discusses the implementation of the pitch extractor.

2.8.3 A Novel Pitch Extractor

The last few sections have described the prediction residual as the result from the linear prediction analysis. It has been shown that the prediction residual contains much information needed for extracting pitch. It is a simple problem to pick appropriate peaks to extract the pitch. It is this problem of pitch extraction that has interested many authors recently.

Examining Figure 13, a repetitive waveform is seen at the period called the pitch period. Note that the waveform has a noisiness which implies a flat amplitude spectrum. It should be noted that for voiced sounds there are other peaks that are also repetitive. These are evidence of the formant frequencies. They are somewhat dampened; however, this is to be expected since the linear predictive filter has the characteristic of spectrally flattening the signal.

It is well known for

$$x(t) = A \cos (2\pi f_o) \tag{2.44}$$

where A is the constant maximum amplitude of the signal, and f_o is the fundamental frequency, the spectrum is

$$X(f) = \frac{A}{2} \delta(f-f_o) + \frac{A}{2} \delta(f+f_o) \tag{2.45}$$

It can be said that the speech waveform is a combination of sinusoids of the type given in (2.44) summed together in a quasi-periodic fashion. The residual signal, $e_f(n)$, can be described in a similar fashion. Therefore, it follows that the spectrum of $e_f(n)$ has impulses at the fundamental and its harmonics identified here by f_o, f_1, ..., f_n. The maximum amplitude is centered at f_o, the fundamental frequency [75]. The higher frequencies are all harmonics, or multiples of f_o. An a priori estimate of f_o for a speech sound can be found using the residual as input. If the spectrum is available, then the frequency of the maximum amplitude determines an estimate of f_o. This estimate is found to be relatively accurate for speech and the prediction residual. In the following, a procedure for extracting the fundamental is given.

The initial step is square the residual. This has a dual benefit in addition to making all calculations positive. First, it makes large quantities larger and second, any small or noise-like quantities are made smaller. The new data corresponding to the set of squared samples are placed in frames of 256 samples each.

Following the initial step, the original sample rate is used to determine the time difference between maximums. It is assumed that the maximums mark the beginning of a new pitch period as set by a threshold.

The threshold is used to select the next maximum. The data set is passed through a peak picker. The peak picker uses the threshold to determine the next peak (maximum). The time between the two peaks is calculated by a differencer function. The system is ready to set a pitch value from the time between the two peaks.

At this point, the a priori estimate of the pitch and the pitch value from the differences are averaged. An error check is made for erroneous pitch values. The error check compares a range of pitch from a low to a high value. Should the averaged value be less or more than a set threshold, an update is sent to recalculate the last averaged pitch in the frame. The process is continued until the end of the frame where the pitch is set. The procedure for estimating pitch is shown in Figure 14. The next section discusses the results in using the pitch extractor.

2.8.4 Pitch Extraction Results

The PITCH program was applied to 39 phonemes, including 16 vowels and diphthongs and 23 consonants. Each sound was held from one-quarter second to one second by a male speaker at normal intensity. Recordings were made on a SONY Model TC-106A tape recorder under anechoic conditions. The sounds were low-pass filtered by a Butterworth filter with a cutoff frequency of 3600 Hertz and samples at 8000 Hertz with nine quantization bits and one sign bit. The computer system quantization level setting was ±10 volts. This gave a quantization level of 20 millivolts. The digitized sound was sampled for 1.5 seconds using 12000 data points for storage. With a limited computer system memory, the beginning of the sound was found and 4096 points were saved. The sound was stored for later use and labeled with an appropriate name. Due to the processing

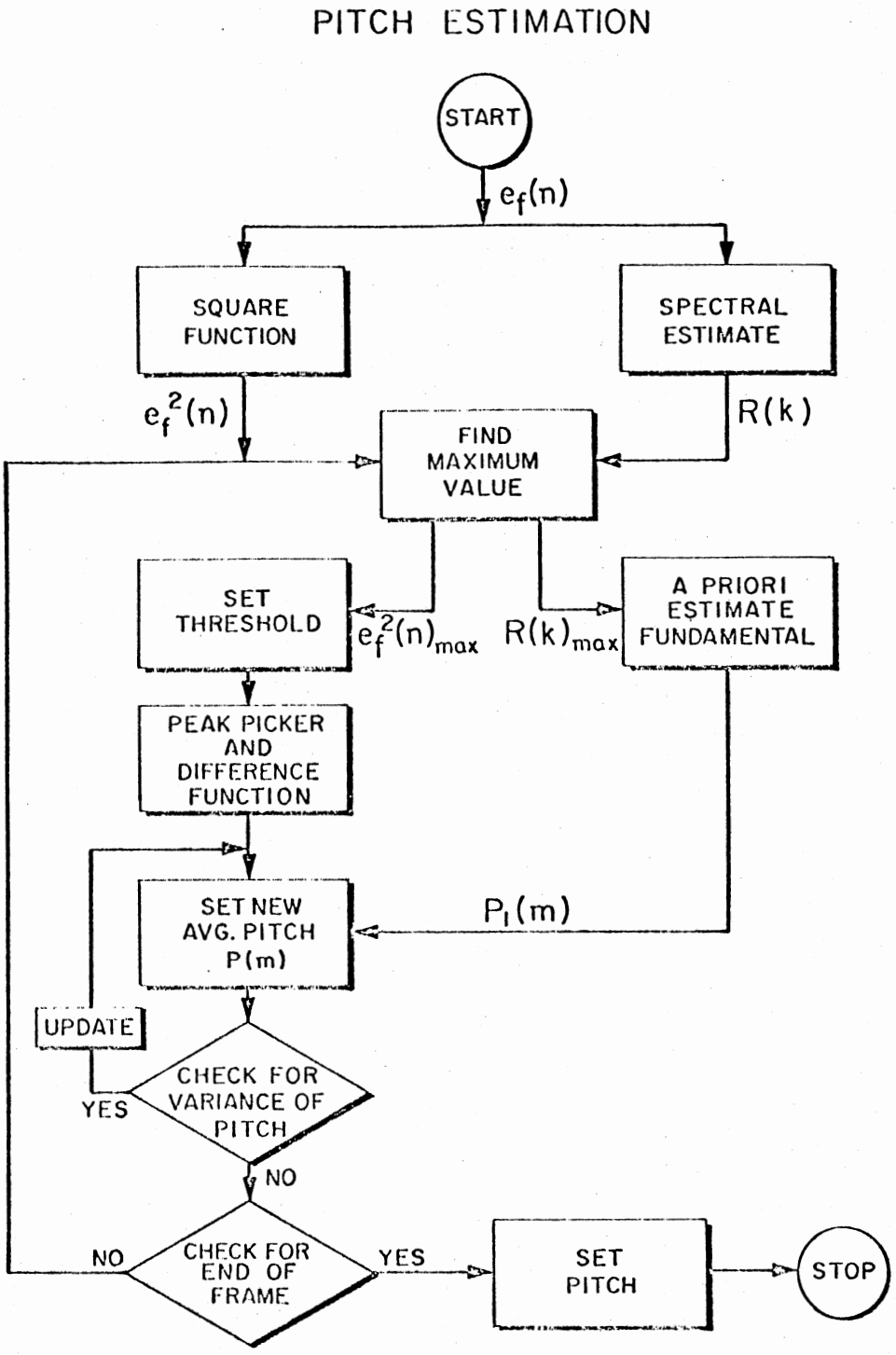


Figure 14. Block Diagram of Pitch Extraction Method

of the INTERDATA 70 computer system, all processing in program PITCH is done in 256-point blocks.

The unified recursive solution to solve Equation (2.10a) is by program AUTO [7]. The efficient resursive solution was discussed in an earlier section. The program INVERS processes the data. The residual data is stored for use in the program FFTMGR. The computation of the spectrum is performed in this program. Spectral values are stored for use in PITCH.

Samples can be plotted for any segment of the sound to aid in visual determination of the pitch period. An example is shown in Figure 13. The results show that the method presented here is an adequate and accurate method for determining pitch period. It is compared to Peterson and Barney's data [111]. Table III gives a comparison between this data and the results obtained from this method. From this table, it can be seen that the results are good. The voiced/unvoiced decision is not a product of PITCH. The FFTMGR routine produces an energy level for the determination of voicing. Voicing errors were made 25 percent of the time. This is due to the fact that the threshold is set to a low level. However, the error check will restrict any wide variance of pitch. If the calculated fundamental frequency is larger than a set threshold value of 400 Hertz, then the corresponding sound is considered as an unvoiced sound and no further calculations are made. These two checks allow for accurate measure of pitch and voiced/unvoiced decision.

Error-free pitch estimation is critical to the overall performance of any low-bit rate coding system. Coding systems that incorporate the residual signal for estimation of the pitch are accurate and adequate. Accuracy can be enhanced by using the residual in a minimum noise

environment. The residual signal formulates a true pitch per frame that can be used at low, synthetic transmission rates. There are several reasons that have been discussed to show that this is a novel approach to pitch extraction and is summarized below. First, it is a two-stage method that estimates the residual spectrum and uses time samples of the residual to calculate the approximation of the pitch. Second, the calculation is done by a thresholding technique which uses the square of samples. Finally, the extraction of the pitch includes an error check that estimates wide variances of the pitch during each frame. From these, it can be seen that this method can be considered as a hybrid technique.

TABLE III

COMPARISON OF FUNDAMENTAL FREQUENCIES

Phoneme	Fundamental from Peterson-Barney [111]	Frequency (Male) from Proposed Method of Pitch Extraction
/i/	136	129
/I/	135	130
/ɛ/	130	125
/æ/	127	135
/ɑ/	124	123
/ɔ/	129	135
/ʊ/	137	126
/U/	141	151
/ʌ/	130	123
/ɝ/	133	140

2.9 Summary

In this chapter, the characteristics of the prediction residual were presented. There is a parallel between the glottal waveform and the residual. The mechanism of speech production and a model of the vocal tract is discussed. Short-time spectral analysis is presented. A review of linear prediction analysis is discussed. A description of the implementation of operations for the calculation of the prediction residual is discussed. A novel approach to pitch extraction is presented.

CHAPTER III

SUB-BAND CODING OF THE PREDICTION RESIDUAL

3.1 Introduction

The average rate that speech is conveyed between humans is about ten phonemes per second. It has been shown that the information rate of speech does not exceed 60 bits/second [2] [67]. For the information content to be preserved, the human must be able to extract the representation of the speech signal at this rate. It is important that the speech is intelligible to the listener, and this aspect is the fundamental consideration of coding speech.

There are two concerns in coding speech signals. First, the message content of the speech must be preserved. The content includes linguistic rules to form thoughts for humans to communicate. Second, the speech signal should be represented so that it can be transmitted. At the receiver, the signal should contain the message without serious degradations.

The interest in speech coding has led researchers to consider techniques that enhance signal quality, reduce transmission rate and cost, without considering the complexity of the coding algorithm [64]. The principle is to enhance the perceptual aspects of speech through the coding method. In this chapter, some basic ideas associated with speech coding are discussed. These include transform coding and sub-band-coding.

Since the speech sounds are characteristically different than most acoustic sounds, it is necessary to consider the properties that include the formants and energy of phonemes. Perceptual aspects that contribute transitional cues for humans to discriminate and differentiate speech sounds are discussed in this chapter. It is known that when human listeners are exposed to speech, available to them are a set of responses that are highly over-learned [65]. The minimum discrimination necessary for absolute differentiation of speech sounds is discussed. Recently, speech coding techniques have contributed efficient methods to enhance the coding of speech signals with few degradations. This chapter discusses some of these methods in Section 3.2. Section 3.3 presents a discussion of the transform coding. Section 3.4 presents the method of sub-band coding in detail. Section 3.5 discusses the determination of frequency bands by the Articulation Index. Section 3.6 presents aspects associated with transitional cueing information for the preception of certain phonemes. Section 3.7 discusses perception of intelligible speech. Section 3.8 describes the basis for coding the prediction residual at the rate of 9600 bits/second.

3.2 Coding Methods

The oldest form of speech coding device is the channel vocoder invented by Dudley [78]. Each of the channels has center frequency ω_k. For each of the channels, the time-dependent Fourier transform is represented as a cosine wave with center frequency ω_k which is phase and amplitude modulated corresponding to the magnitude and phase angle, respectively of each transform. Therefore, each channel is thought of as a bandpass

filter with center frequency ω_k and impulse response w(n). This is shown in Figure 15.

The analysis section consists of a bank of channels as in Figure 15 with frequencies distributed across the speech band. Figure 16 shows a complete channel vocoder analyzer.

The basic diagram for the synthesizer is somewhat different. The specific channel controls the amplitude of its contribution to a particular channel; while the excitation signals control the detailed structure of the output of a given channel. The voiced/unvoiced decision serves to select the appropriate excitation generator, i.e., random noise for unvoiced speech and pulse generator for voiced speech. A block diagram is shown for the synthesizer in Figure 17. Channel vocoders operate in the range of 1200 bits/second to 9600 bits/second. They are also referred to as source coders and produce speech of a synthetic nature when at bit rates below 4800 bits/second.

A major contribution of a channel vocoder is the reduction in bit rate; however, direct representation of the pitch and voicing information is not achieved. Therefore, this can be considered as a weakness.

The LPC vocoder is a very important application of linear predictive analysis in the area of low bit rate encoding of speech. It is shown in Figure 18.

The basic LPC analysis parameters consists of a set of P predictor coefficients, the pitch period, a voiced/unvoiced parameter and a gain parameter. The vocoder consists of a transmitter which performs the LPC analysis and pitch detection. These parameters are coded and transmitted. They are decoded and synthesized to output speech. This category of

CHANNEL VOCODER COMPONENT

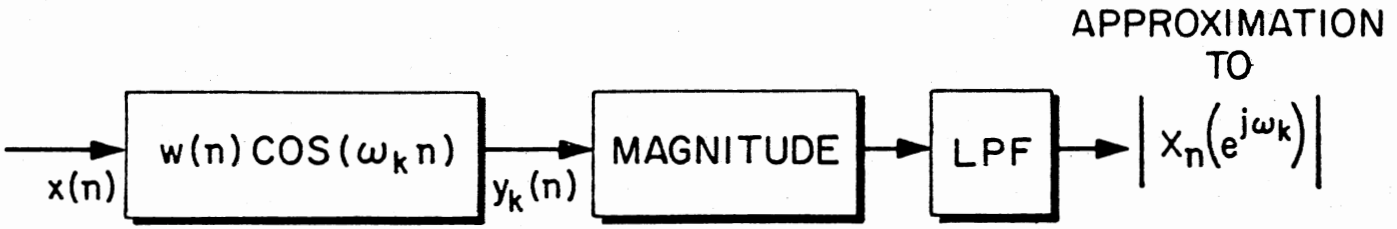


Figure 15. Basic Component of Channel Vocoder

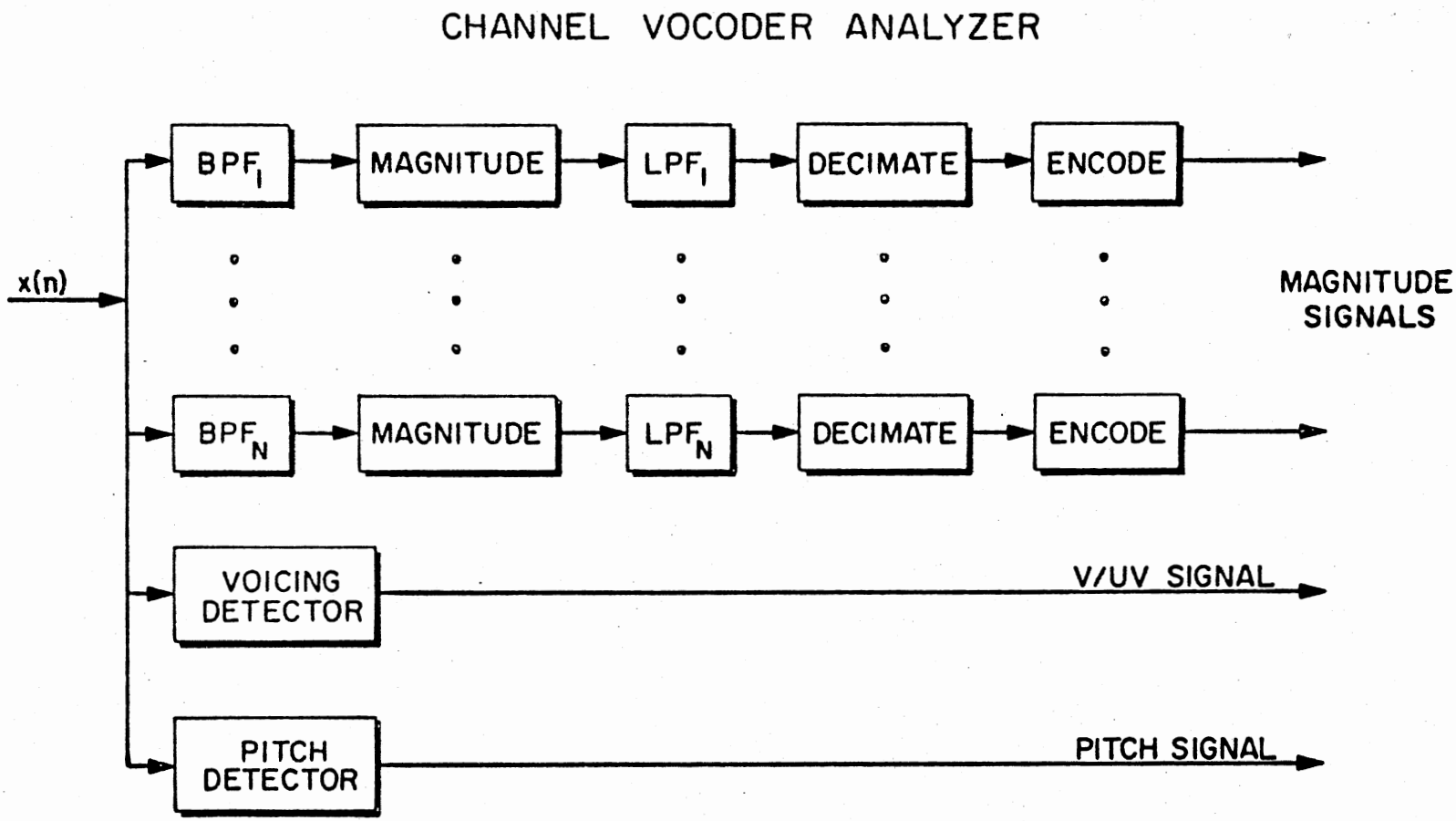


Figure 16. Block Diagram of Channel Vocoder Analyzer

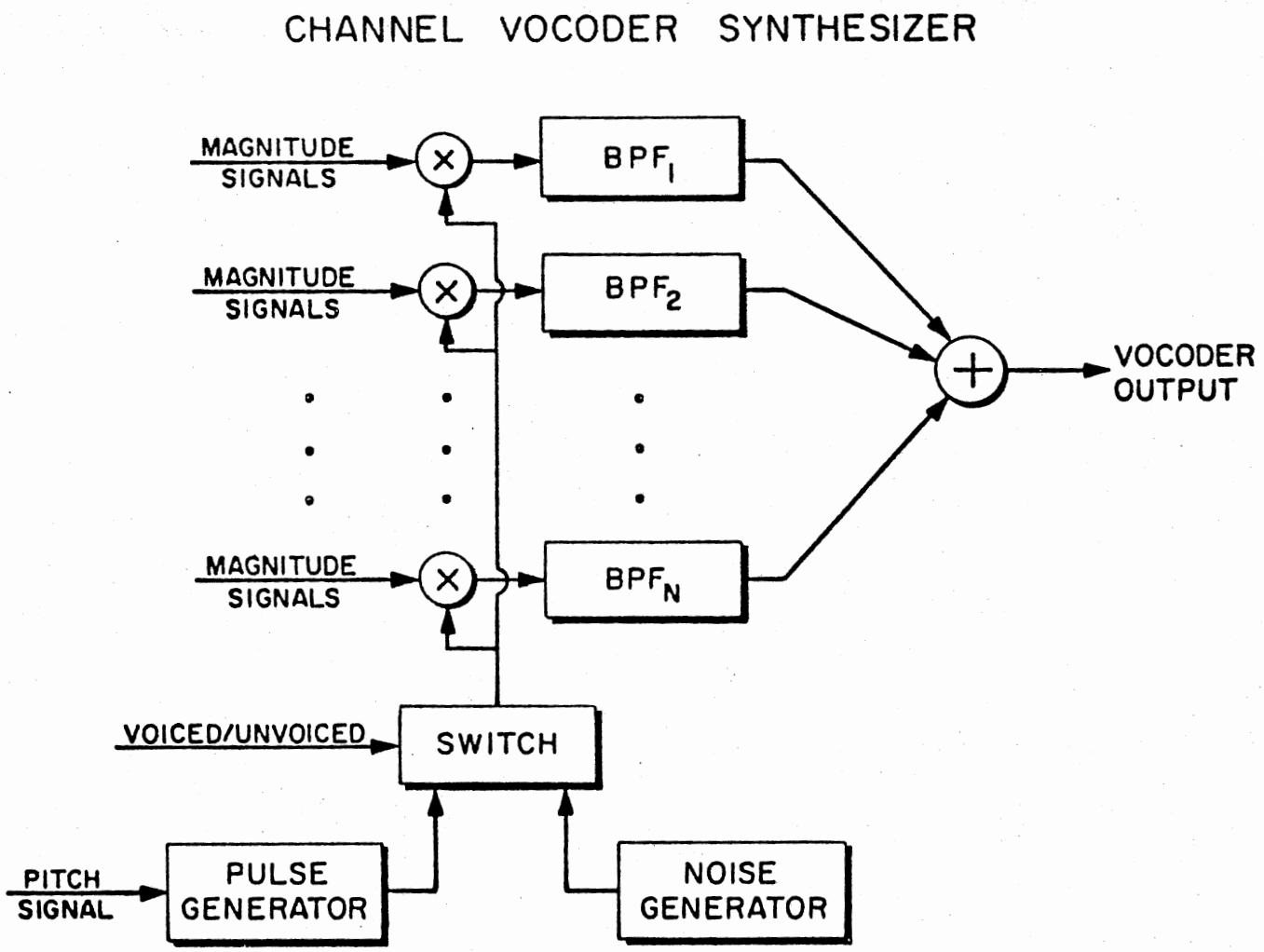


Figure 17. Block Diagram of Channel Vocoder Synthesizer

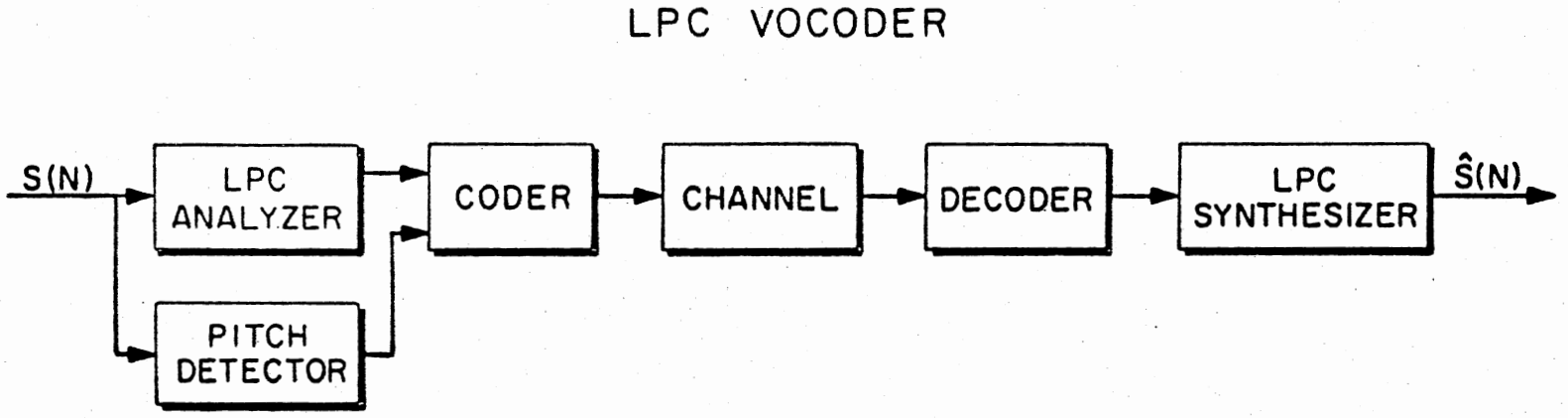


Figure 18. LPC Vocoder

coding is of the vocoder type. In the following, the discussion of various aspects is included.

Speech coding can be divided into two broad categories, waveform coders and vocoders [79]. There has been some metion of a few types of vocoders earlier. Waveform coders generally attempt to reproduce the original speech waveform according to some fidelity criteria. On the other hand, vocoders model the input speech according to some speech production model; then, synthesize the speech from the model. The basic make-up for coding the prediction residual in this thesis is of a vocoder model. However, techniques of waveform coding are also used. It has been shown that waveform coders tend to give better quality speech that is robust; whereas, vocoders tend to be more synthetic [64] [79]. Borrowing from the techniques of efficient waveform coders, it is conceivable to define an acceptable coding algorithm to meet quality standards at low-bit rates of transmission. A primary interest has been to produce the transmitted speech with the minimum bit rate and still meet acceptable quality [80]. Previously mentioned were methods available to date for coding of the residual. Efficient methods to improve the coding techniques are presented for coding the prediction residual.

It has been recognized that there are two efficient methods of waveform coding [79]. These are: (1) transform coding (TC) [81] and (2) sub-band coding (SBC) [36] [37]. These are characterized as frequency-domain coders, whereas examples of PCM, differential PCM, and DM are the time-domain coders. Frequency-domain coders are perceptually better than time-domain coders because they tend to exploit the pitch of the speech waveform for bit rates below 16000 bits/second. They tend to look at the spectrum of speech in blocks, whereas the predictive systems look at

adjacent samples. These two methods will be explained in detail in the next two sections.

3.3 Transform Coding

With transform coding (TC) [81], the system of speech samples is grouped into blocks, where each block corresponds to the windowed segment of the speech signal. These blocks of speech are transformed into a set of transform coefficients; then, the coefficients are quantized independently and transmitted. An inverse transform is taken at the receiver to obtain the corresponding block of reconstructed samples of speech (see Figure 19).

A basic assumption in this method is that the speech source is stationary and has a variance of σ^2. The successive source output samples are arranged into the N-vector X; this vector X is linearly transformed using a unitary matrix A, i.e.,

$$Y = AX \tag{3.1}$$

where A, in general, is complex, and

$$AA^* = I \tag{3.2}$$

where * denotes the transpose conjugate. The elements of $\hat{Y}$ are the transform coefficients. These are independently quantized, yielding, $\hat{Y}$. The vector $\hat{Y}$ is transmitted to the receiver and then inverse transformed. Then

$$\hat{X} = A^* \hat{Y} \tag{3.3}$$

Since the vector $\hat{X}$ is reconstructed output, distortion is involved. For unitary matrices the averaged mean-squared overall distortion of the

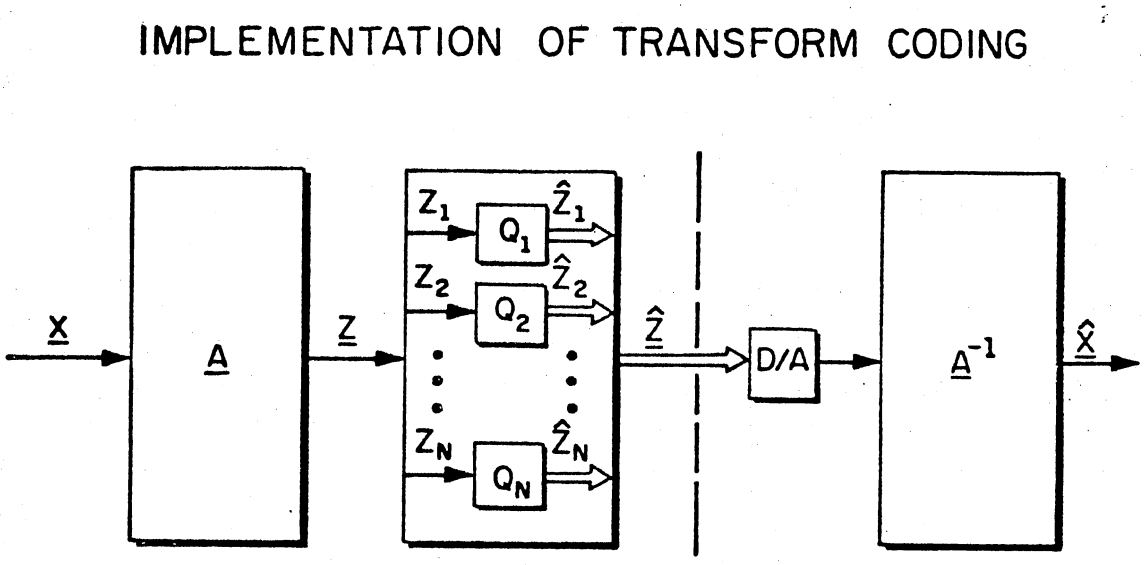


Figure 19. Block Diagram of the Implementation of Transform Coding (After Zelenski and Noll, 1977).

transform coder is equal to the quantization error [82]

$$\bar{D} = \frac{1}{N} \cdot E\{(X - \hat{X})^T \cdot (X - \hat{X})\}$$

$$= \frac{1}{N} \cdot E\{(Y - \hat{Y})^T \cdot (Y - \hat{Y})\} \qquad (3.4)$$

where $E\{\ \}$ represents the expectation. The minimization of $\bar{D}$ will yield an optimum bit-assignment rule and an optimum transform matrix A [81]. Let J_i be the number of bits/sample needed for the coefficient Y_i (an entry in the Y vector) of variance σ_i^2 so that the mean-squared distortion $D_i = E[Y_i - \hat{Y}_i)^2]$ is not exceeded. Then [82]

$$J_i = \delta + \frac{1}{2} \log_2 \left[\frac{\sigma_i^2}{D_i} \right] \qquad (3.5)$$

where δ is a correction factor that takes into account the performance of a practical quantizer. The optimum number of bits for the quantizer can be obtained by minimizing the average distortion

$$\bar{D} = \frac{1}{N} \sum_{i=1}^{N} D_i \qquad (3.6)$$

with the constraint of a given average bit rate

$$\bar{R} = \frac{1}{N} \sum_{i=1}^{N} J_i = \text{constant} \qquad (3.7)$$

The optimum bit assignment is [81]

$$J_i = \bar{R} + \frac{1}{2} \cdot \log_2 \frac{\sigma_i^2}{\left[\prod_{i=1}^{N} \sigma_i \right]^{1/N}} \text{ bit/sample} \qquad i = 1, 2, \ldots, N$$

(3.8)

The average distortion is found to be

$$\overline{D} = 2^{2\delta} \cdot 2^{-2\overline{R}} \cdot \left[\prod_{j=1}^{N} \sigma^2 \right]^{1/N} \tag{3.9}$$

Here the distortion introduced by the transform coding scheme depends on the distribution of variances. In addition, $\overline{D}$ is found to be the geometric mean of the variance. This leads to the solution of A matrix. Let R_{xx} and R_{yy} be the covariance matrices of X and Y, then

$$\det R_{yy} \leq \prod_{i=1}^{N} \sigma_i^2 \tag{3.10}$$

and for any unitary matrix A

$$\det R_{xx} = \det R_{yy} \tag{3.11}$$

In particulre, the variances σ_i^2 are along the diagonal of R_{yy}; then,

$$\det R_{xx} = \prod_{i=1}^{N} \lambda_i \tag{3.12}$$

where λ_i are the eigenvalues of R_{xx}. Therefore, the minimum distortion is found if the variances, σ_i^2, are equal to the eigenvalues of R_{xx} [81]. The Karhunen-Loève transform (KLT) has the property that $\sigma_i^2 = \lambda_i$ for all i.

Other unique properties of KLT are: (1) transform coefficients are uncorrelated, (2) the covariance in the KLT domain is diagonal, and therefore, the transform coefficients can be quantized independently without the loss of performance [83].

It has been noted that the KLT gives optimum performance; however, there is a lack of a fast algorithm for the computation of the coefficients. In addition, the computation is quite complex. Since speech is a quasi-periodic signal, transform coding would not be efficient unless

adaptive methods are used. However, this area still needs additional studies. Zelenski and Holl presented promising results. Tribolet and others [38] have done additional work in this area also. Zelenski and Noll experimented with the Walsh-Hadamard transform (WHT), the discrete slant transform (DST), the discrete Fourier transform (DFT), and the discrete cosine transform (DCT) to compare with the KLT. All these have fast algorithms and are signal independent. Zelenski and Noll found that the basis vectors of the DCT and KLT are close; however, the KLT is signal dependent. It has been shown that the performances of the DCT and KLT are similar [84]. The studies of Tribolet and others found TC to be complex and costly; however, this method proves to be superior when compared to other systems [38].

3.4 Sub-Band Coding

It is desired to retain the basic components of speech composition and phonemic quality. TC is a very efficient method of completing the endeavor; however, due to cost and complexity, it was discarded. The method of sub-band coding [36] has some very distinct advantages whereby the original goal can be met in order to secure as much of the speech signal as possible. One criterion, perceptual in nature, is the retention of transitional information. Also, the intelligibility of speech can be maximized by the use of the Articulation Index [29], which is discussed in Appendix C.

With sub-band coding the frequency spectrum is partitioned such that each sub-band contributes accordingly to the speech intelligibility which is quantified by the Articulation Index. The Articulation Index is a weighted fraction representing, for a given speech channel and noise

condition, the effective proportion of the normal speech signal which is available to a listener for conveying speech intelligibility [86]. The speech spectrum can be divided into 20 frequency bands contributing 5 percent each to the Articulation Index. In this case, the frequency spectrum can be bandpass filtered in such a way that they contribute equally to the Articulation Index. An example given by Crochiere and others [36] in Table IV addresses a sub-band partitioning of four bands between 200 to 3200 Hz.

TABLE IV

SUB-BAND PARTITIONING EXAMPLE

Sub-Band No.	Frequency Range (Hz)
1	200 - 700
2	700 - 1310
3	1310 - 2020
4	2020 - 3200

Obviously, there are other possibilities of partitioning the speech band [37]. Each band contributes an equal 20 percent to the total Articulation Index. The total Articulation Index is 80 percent, which corresponds to a word intelligibility of approximately 93 percent [36].

Sub-band coding has another advantage which involves quantization. Each sub-band is quantized separately and each band contains its own distortion and, therefore, quantization noise could be considered separately for each band [36]. Furthermore, because of the nature of the spectrum of speech, the detectability of this distortion is not the same at all frequencies.

Since the proposed method is based upon sub-band coding the residual, the presentation is in terms of the prediction residual, $e_f(k)$.

For the following discussion, assume that the sub-bands are partitioned as shown in Figure 20. Let the width of each of these bands be identified by

$$W_n = \omega_{n+1} - \omega_n \qquad n = 1, 2, \ldots, N=4 \tag{3.13}$$

where ω_n corresponds to the edges of these bands. The implementation of the sequence of operations leading from the residual to the coded output for transmission is shown in Figure 21. Also, shown in Figure 21 is the implementation at the receiver. From this figure, it follows that

$$r(k) = (e_{fn}(k) \cos \omega_n k) * h_n(k) \tag{3.14}$$

where $e_{fn}(k)$ corresponds to the output of the nth bandpass filter and $h_n(k)$ corresponds to the impulse response of the nth lowpass filter. It is clear that

$$W_n \leq 2\omega_n \tag{3.15}$$

in order that the frequency bands are properly separated. Then $r(k)$ is decimated to the rate $2W_n$ from the original sampling frequency. This signal is then encoded and multiplexed with the other channels. At the receiver, the signal is demultiplexed, decoded, interpolated, demodulated

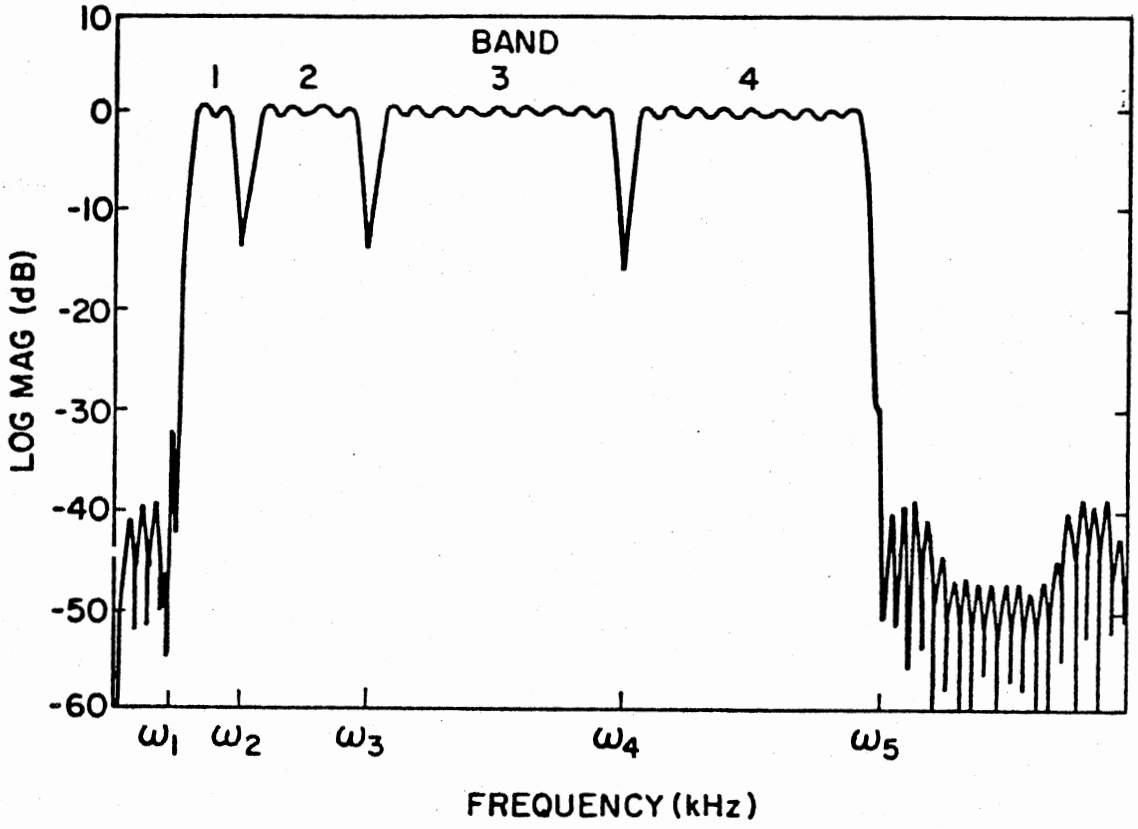


Figure 20. Partitioning of Frequency Spectrum into Four Sub-Bands (After Crochiere, 1976)

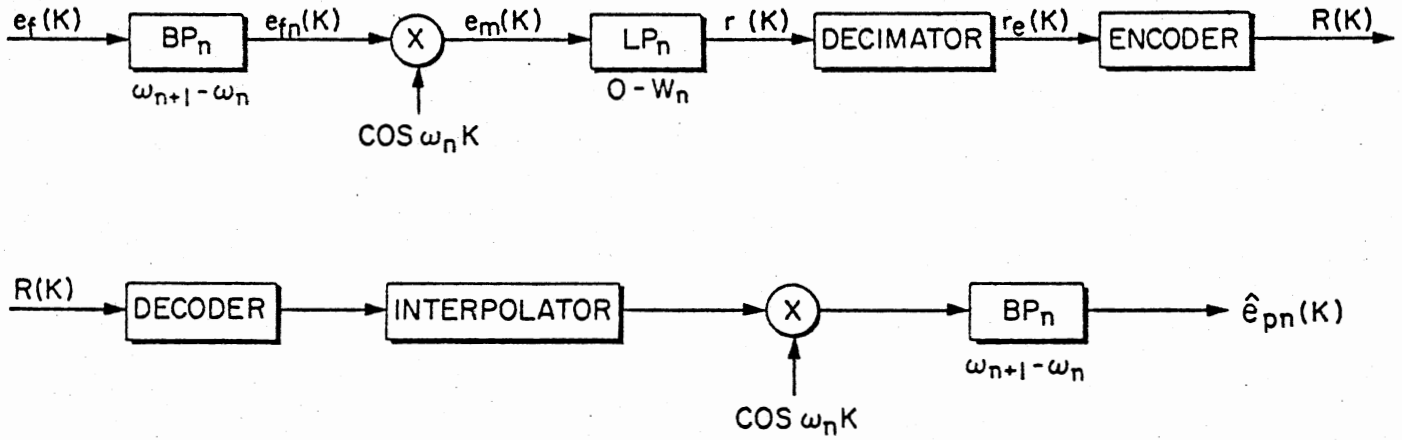


Figure 21. Sequence of Operations Relating to the nth Sub-Band

and bandpass filtered to give $\hat{e}_{fn}(k)$. This is shown in Figure 21. The nth sub-band is then summed with other bands to produce $\hat{e}_f(k)$, which is the sub-band coded and decoded version of the signal. The total implementation of the system will be discussed later.

3.4.1 Sub-Band Coding and Transform Coding

Earlier it was pointed out that frequency-domain coders can be considered as a good basis for an efficient coder. In this section the relationship between sub-band coding and transform coding is discussed.

Considering the ideal case, in which there are M sub-bands corresponding to the M samples, let the discrete cosine transform (DCT) of the residual signal, $e_f(k)$, $k = 0, \ldots, M-1$, be represented by [84]

$$\left.\begin{aligned} \alpha_o &= \frac{1}{\sqrt{M}} \sum_{k=0}^{M-1} e_f(k) \\ \alpha_n &= \frac{\sqrt{2}}{\sqrt{M}} \sum_{k=0}^{M-1} e_f(k) \cos \frac{(2k+1)n\pi}{2M} \end{aligned}\right\} \quad n = 1, \ldots, M-1 \qquad (3.16)$$

Correspondingly, the residual signal $e_f(k)$ is given by

$$e_f(k) = \frac{1}{\sqrt{M}}\alpha_o + \sqrt{\frac{2}{M}} \sum_{n=1}^{M-1} \alpha_n \cos \frac{(2k+1)n\pi}{2M} \qquad k = 0, 1, \ldots, M-1 \qquad (3.17)$$

which obviously corresponds to the inverse discrete cosine transform (IDCT). Using

$$\omega_n = \frac{(2k+1)n\pi}{2M} \qquad (3.18)$$

in Figure 21, it is seen after modulation and low-pass filtering

$$r(k) = \alpha_n \qquad k = 0, 1, \ldots, M-1 \qquad (3.19)$$

Since there are M sub-bands corresponding to the M frequencies, and since r(k) is a constant, it follows that after the decimation only one point is given for each band and that value is α_n. The encoder in Figure 21 codes the DCT coefficient. This points out the fact that in the ideal case (i.e., filters and modulators are ideal), the sub-band coding will be equivalent to the discrete cosine transform coding. Obviously, the discussion above can be generalized for the case wherein there are N sub-bands ($N \leq M$) rather than M.

It is clear that where the components in the sub-band coder are non-ideal, the r(k) are not equal to α_n. Further work is necessary in quantifying the difference between r(k) and α_n [85].

Noting the simplicity in the sub-band coder and also noting the relationship between the transform coder and sub-band coder, the sub-band coder is more practical.

3.5 Determination of Frequency Sub-Bands Based on Articulation Index

The Articulation Index (AI) is a weighted fraction representing, for a given speech channel and noise condition, the effective proportion of the normal speech signal which is available to a listener for conveying speech intelligibility [29].

In this section, the methods of determining how to achieve maximum intelligibility based on using the AI are examined. There are two methods for computing AI. The first method, called the 20-band method by French and Steinberg [86], is based on measurements or estimates of the spectrum of the speech and noise present in each of the 20 continuous bands of frequencies. Each band contributes equally to the speech

intelligibility. The second method, known as the octave-band method, is derived from the first method. It requires measurements of the speech and noise present either in certain one-third-octave-band or in certain full octave bands.

Some researchers consider these two, i.e., one-third-octave-band and full octave-band measurements, as different methods. The octave-band method is not as sensitive to variations in the speech and noise spectra as the 20-band or the one-third-octave-band method. An example where it falls apart is in situations where an appreciable fraction of the energy of the masking noise is concentrated in a band of frequency that is one octave or less in width; under these conditions, the one-third-octave-band or the 20-band method would be better to use.

The 20 frequency bands are those specified by Beranek for male voices [87]. These bands are shown in Table XXIV in Appendix C. In order to use the 20-band method to calculate the AI, the peaks of the spectrum of the speech signal (PSS) must be approximated first. The level depends on if the speech is spoken through earphones or a loudspeaker. There is an adjustment to either case of -65 dB which is considered as the over-all long-term rms sound-pressure level of an idealized speech spectrum. However, with the loudspeaker, an additional amount is adjusted according to Table V [29]. This is due to the assumption that the room is semireverberant; whereas, earphones do not present reverberance.

These corrections are obtained from experiments conducted in a reverberant room using a loudspeaker and from experiments conducted in an anechoic chamber [29].

Also an additional correction must be added to correct for the noise spectrum. This is shown in Table VI [29]. The noise that reaches the

listener's ear is assumed to be that of a steady-state nature. All noises in the listener's environment and the noise in transmission systems are combined to arrive at the noise spectrum level.

TABLE V

ADJUSTMENTS TO THE SPECTRUM OF THE SPEECH SIGNAL

Maximum Spectral Values of Speech Signal	Amount to be Subtracted
85 dB	0 dB
90	2
95	4
100	7
105	11
110	15
115	19
120	23
125	27
130	30

The corrected noise spectrum (NS) has the effect of masking the speech signal. The noise spectrum increases at a faster than normal rate when the band sensation level of the speech sound exceeds 80 dB [86]. This band sensation level is defined as the difference in decibels between

the sound integrated over a frequency band and the sound pressure level of that band when the speech sound is at the threshold of audibility in an anechoic room. The increase in masking is taken into account in the calculation of AI by adding to the PSS. If the band sensation level of the sound exceeds 80 dB at the center frequency of a band, then the PSS is increased by the amount that is shown in Table VI.

TABLE VI

ADJUSTMENTS FOR NOISE SPECTRUM

Band Sensation Level	Added Amount
80	0
85	1
90	2
95	3
100	4
105	5
110	6
115	7
120	8
125	9
130	10
135	11
140	12
145	13
150	14

The noise spectrum level (NS) is compared to PSS at the mid-frequencies of the 20 bands given in Table XXIV in Appendix C. Values that are zero or less are set to zero. When PSS exceeds the noise by 30 dB, then that difference is set to 30. This is due to the limitation on the dynamic range of speech [87].

The Articulation Index is defined as

$$AI = \sum_{n} W_n \cdot (\Delta A)_{max} \tag{3.20}$$

where

$(\Delta A)_{max}$ is the contribution from one band and has a maximum value of 0.05.

W_n is the percent of maximum contribution by any one band

and

$$W_n = \frac{PSS - NS}{30} \tag{3.21}$$

where 30 represents the dynamic range of the speech band and is a normalized so that W_n is limited to unity. Therefore, for 20 bands, the normalization is limited to 600. An illustrative example is given by Kryter [29].

Consider the one-third-octave-band and octabe-band method. The center and cut-off frequencies for these are shown in Tables VII and VIII [29].

With the one-third-octave and octave-band methods, the correction levels shown in Table V should be considered for signals received from the loudspeaker. Also, the NS must be calculated from Table VI, and the weighting factors need to be computed from (3.21) for each band. These

are then summed to give the AI for the speech system operating under the noise conditions and the level of speech.

TABLE VII

FREQUENCIES RELATED TO ONE-THIRD-OCTAVE-BAND METHOD

One-Third-Octave Band	Center Frequency
179 - 224	200
224 - 280	250
280 - 353	315
353 - 448	400
448 - 560	500
560 - 706	630
706 - 896	800
896 - 1120	1000
1120 - 1400	1250
1400 - 1790	1600
1790 - 2240	2000
2240 - 2800	2500
2800 - 3530	3150
3530 - 4480	4000
4480 - 5600	5000

TABLE VIII

FREQUENCIES RELATED TO OCTAVE METHOD

Octave Band	Center Frequency
180 - 355	250
355 - 710	500
710 - 1400	1000
1400 - 2800	2000
2800 - 5000	4000

To consider how the different methods compare for the same speech signal and masking noise, Kryter computed the AI for each of these methods. For 20-band method, AI = 0.38; for one-third-octave method, AI = 0.33; and for octave-band method, AI = 0.28. Since the 20-band method is the basic method from which all others are derived, it provides the "correct" AI and the others are compared to this AI.

The AI can be compared to estimated speech intelligibility scores as shown by graph in Figure 22. It is noted that the intelligibility score is highly dependent on the constraints placed on the message communicated.

The greater constraint (for instance, the smaller the amount of information content in each item of message), the higher the percent intelligibility score for a given AI. No single AI can be used as a criterion for an acceptable communication value. It is a function of messages transmitted and the enunciation of the talker [29].

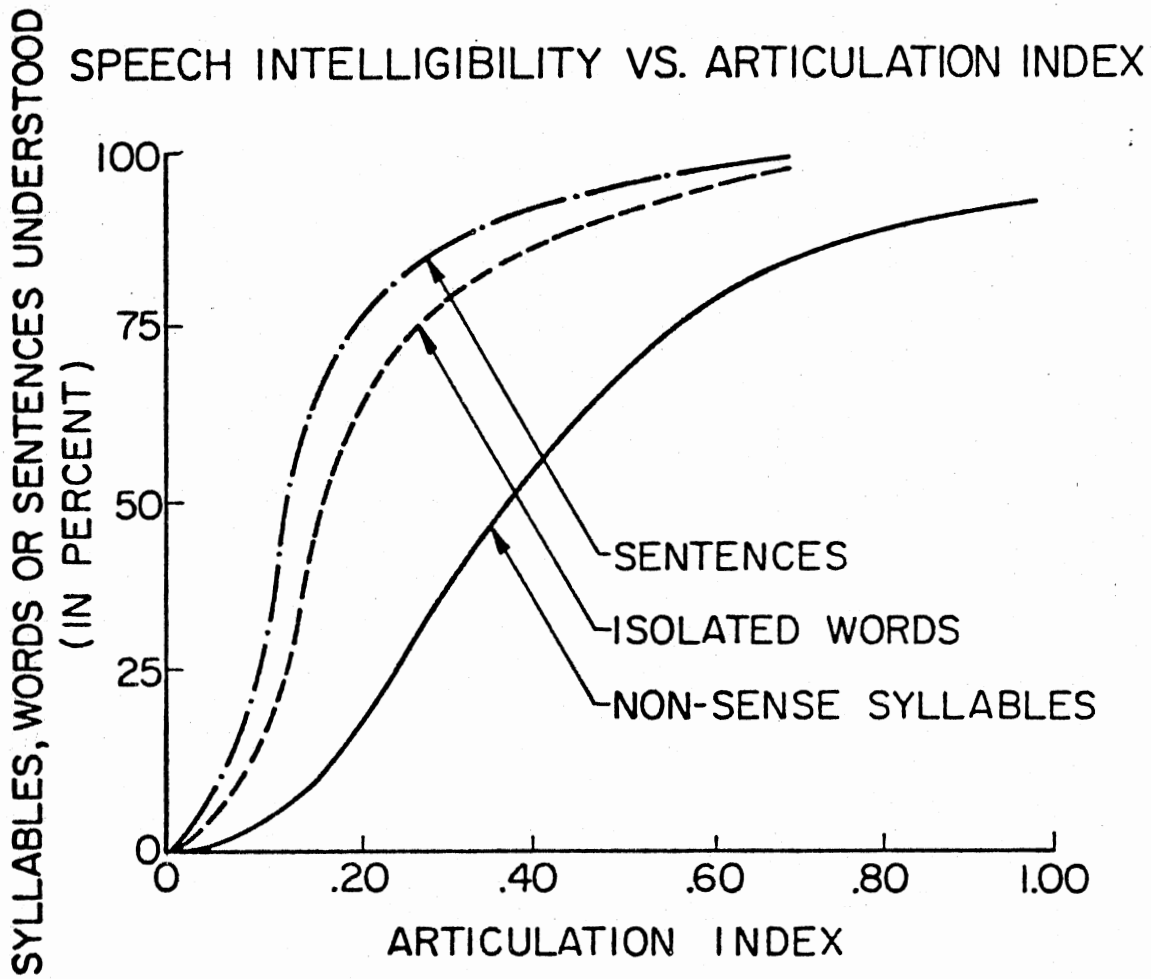


Figure 22. Relation Between AI and Various Measures of Speech Intelligibility (After French and Steinberg, 1947)

The Articulation Index is a good quantitative measure of speech intelligibility. Speech communication can be enhanced by an equal application of AI across the speech spectrum in the sub-band coding. In the next section, the transitional cueing of phonemes, another aid that adds to speech communication is discussed.

3.6 Transitional Information

The human speaks in an uninterrupted and continuous fashion to communicate thoughts. The underlying basis for communication is the phonemic structure that connects itself by means of transitional cues for the perception of certain phonemes [1]. It is the transitional information that must be enhanced to aid the perception needed for absolute discrimination of speech-like sounds [2]. Transitional cues are a set of frequency shifts which occur in the second-formant region where a consonant and a vowel join. The perception of a given phoneme is strongly conditioned by the transitional information of its neighbors [2].

The identification of phonemes has been studied under various conditions by a group at the Haskins Laboratories [1]. Many of their experiments have used synthetic syllables. The combinations of syllables included consonant-vowel (CV) syllables. The consonant is usually a stop out of a group of phonemes with the same voicing. The vowels were maintained at two formants. Further work has been done by Rabiner [88] for synthesis of phonemes by rules. These concluded that one frequency variable of the consonant was generally adequate to distinguish that a consonant of the group was uttered. To further distinguish the

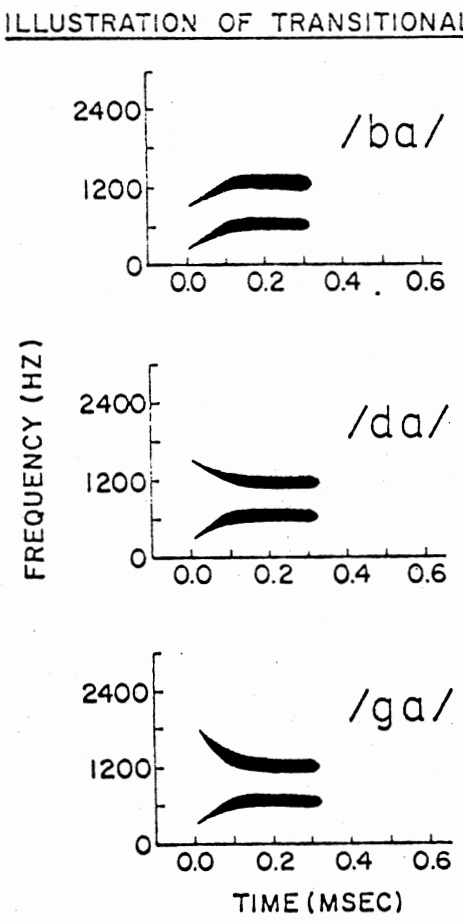
consonant, the stop-vowel formant transitions were necessary to perceive the consonant.

Figures 23(a) and 23(b) illustrate the stop-vowel formant transitions. In Figure 23(a), the vowel /a/ has first and second formants occurring at 700 Hz and 1200 Hz respectively. It is seen that the consonants /b/, /d/ and /g/ demonstrate a different rise or fall in the second formant region. The second formant varies because each consonant has a different place of articulation. The place of articulation for /b/, /d/ and /g/ are front, middle and back, respectively. It is seen in Figure 23(a) that the consonants appear to commence from some trajectory determined by their place of articulation.

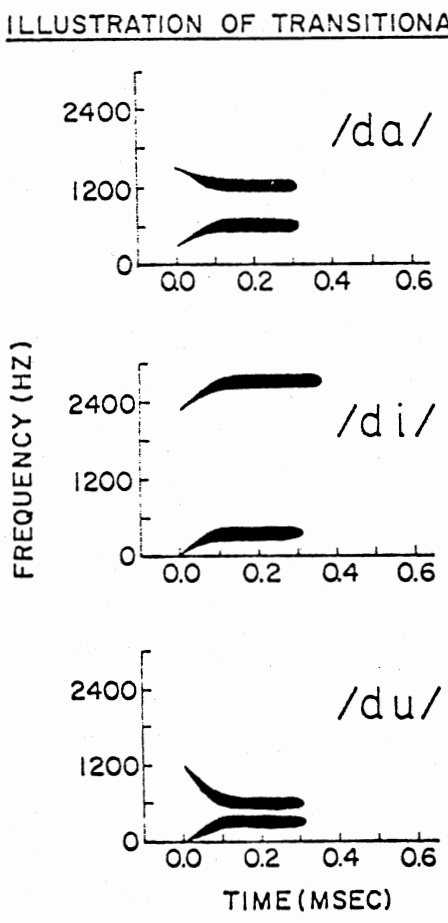
The trajectory point is further illustrated in Figure 23(b). This figure uses the consonant /d/ and three vowels, /a/, /i/ and /u/. It is shown that the consonant /d/ has a loci of points that commence in the region of 1600 Hz for the second formant. It has been shown that consonants exhibit this property of transition from a particular frequency to the steady-state value of the vowels [1].

The consonants that are perceptually heard with falling second formants to the vowel /a/ are /d/ and /g/. The consonant /b/ is heard with a rising second formant to the vowel /a/. It is noted that a shift in second formant frequency is bounded. With falling transitions of the second formant, /g/ is heard for steady-state levels of frequency between 2280 and 3000 Hz; however, between 1320 and 2280 Hz the sound could be /g/ or /d/; and, below 1320 Hz, it is identified as /d/ [1].

The importance of second-formant transitions is shown for perceptual purposes. Differences in the acoustic speech signal are due to the excitation and vocal tract configuration for different consonants. These



a.) Phoneme /a/



b.) Phoneme /d/

Figure 23. Transitional Cueing for Consonant-Vowel for Phonemes /a/ and /d/

will differ as shown in Figure 23 by the transition regions. The consonant transitions are the principal cues for the perception of a particular consonant. The transition occurs because the vocal tract has one shape for the vowel and one shape for the consonant. The change in the vocal tract and the effect that the glottal pulse has on vowels has been addressed recently [12]. Later, the coding aspects of transitional information will be discussed.

3.7 Relation of Perception to Intelligible Speech

A topic that has been mentioned several times before is perception. Perception related to the Articulation Index and transitional information together for discrimination of speech sounds. A quantitative description of speech perception is not possible. However, in a qualitative sense, speech perception can be enhanced when the intelligibility of speech is increased. In this section, several aspects of speech perception will be discussed to show the need to address this subject.

Speech perception can be defined as the ability for humans to discriminate and differentiate the character of speech sounds. Discrimination is examined along fundamental dimensions of the hearing mechanism and, in general, one dimension at a time. The ear takes measurements and makes differential comparisons. These comparisons may be of frequency and intensity. The over-learned senses of the brain distinguishes the speech from other periodic waves. Further, the speech must be broken in to its discrete elements, the phonemes. Once the signal is perceived as speech, there are other factors that determine the fundamental characteristics of recognizing intelligible speech.

The ability to recognize and understand speech determines intelligibility. The intelligibility of speech may be affected in several ways [86]. These may include echoes, phase distortion, or reverberation. Unnatural sounding speech can influence intelligible understanding of speech sounds. The intensity of the speech may affect intelligibility of speech received by the ear. Noise in a transmission medium may affect intelligibility by masking the speech. The talker and listener have several factors that can cause unacceptable intelligibility related to the speech [86]. These are given below:

a. The basic characteristics of the speech can be destroyed.

b. The electrical and acoustic instruments which operate between the talker and the listener may not be adequate.

c. The condition under which the communication takes place may not be acceptable.

d. As a result of c., the behavior of the talker and listener may be modified by the characteristics of the communication system.

The perception of intelligible speech is related to the amount of information spoken. This is shown in Figure 22. The exactness with which the listener identifies speech sounds is related to the size of the vocabulary and the sequence or context of the message. As seen from Figure 22, the more predictable the message is, the better the intelligibility. It has been shown that as the vocabulary size increases, a higher signal-to-noise ratio is necessary to maintain performance [2].

Perceptual aspects of speech are influenced greatly by semantics and context. The ability to predict the speech utterance enhances intelligibility. The grammatical rules of a language are part of the human over-learned senses [65]. Consequently, the language prescribes a

certain allowable sequence of words. The semantic factors occur as part of the rules because certain words must be associated with meaningful units [66]. It has been shown that intelligibility of speech is substantially higher when a grammatically correct and meaningful sentence is spoken than when using the same words randomly [65]. The over-learned senses reduce the number of alternative words from the context, and therefore, the listener has improved intelligibility.

The application of speech perception is an adaptive process. The listener uses the detection procedure within the reception system of the ear to determine the speech communication process. The listener can absolutely identify speech when given the basic sound elements of the speech. The sound elements are discriminated and differentiated from other periodic sounds to perceive speech. If the speech is intelligible, the exactness is not only related to how good the transmission medium is but also to the length of the utterance and its context. These concepts are applied in the next section to aggregate a coding algorithm for transmission of perceptually enhanced speech.

3.8 Basis of Coding the Predictional Residual

A coding method is presented to perceptually enhance the speech. The method uses sub-band coding (SBC) for coding the prediction residual. Besides SBC being conceptually simple, it has the additional advantage that each sub-band is quantized separately and each band contains its own distortion. It should be pointed out that the input to the sub-band coder is the residual signal rather than the speech signal. Some of the reasons for this approach are:

a. A more efficient bit distribution based on energy/frame.

b. A more pronounced pitch information in the residual signal, and

c. An ideal input for the synthesizer at the receiver.

In an earlier section, the advantages of using the Articulation Index in SBC have been discussed. Each sub-band is selected such that each contributes equally to the Articulation Index [36]. However, it has been shown that "satisfactory" performance can be expected if this equal contribution to the articulation criterion can be met within a factor of two [37] [87]. This relaxation of the criteria was allowed for integer-band sampling with good results [36] [37]. That is, the sub-bands are between $m_i\ \omega_i$ and $(m_i + 1)\ \omega_i$, where m_i is an integer. The method has popularity because it eliminates the need for modulators. Even though the integer-sampling method requires less hardware, the selection of sub-bands using the articulation criteria would give better perception. There has been some research done in the selection of the sub-bands by this method [37]. Also, it should be pointed out that the sub-band selection depends on the multiplexing of the encoded speech [37]. This subject will be further discussed in Chapter IV.

The coding scheme of the residual is based on enhanced transitional cues. It has been shown that the second formant is important for perceptual purposes. The exact development will be discussed in this section.

The spectrum of the signal is used for calculation of the energy. The energy can be represented by [108]

$$E \cong \frac{1}{N} \sum_{k=0}^{N-1} |E_f(k)|^2 \qquad (3.22)$$

where $E_f(k)$ corresponds to the discrete Fourier transform (DFT)

coefficients of the signal $e_f(k)$, which can be computed by using the fast Fourier transform (FFT) algorithm.

Equation (3.22) is applied to the prediction residual to compute the energy. The spectrum of the prediction residual is partitioned into four sub-bands as stated before. Using (3.22), the energies in each sub-band can be expressed by

$$E_n = \frac{1}{N} \sum_{k=0}^{N-1} |E_{fn}(k)|^2 \qquad n = 1, 2, 3, 4 \qquad (3.23$$

where $E_{fn}(k)$ is the DFT coefficient of the signal corresponding to the nth sub-band.

Now the total energy can be expressed by

$$E_T = \sum_{n=1}^{4} E_n \qquad (3.24)$$

Among speech sounds, E_T has wide variance. Previous researchers have not studied the variations in E_T of the speech sounds for each prediction residual. This aspect is discussed in the next section.

3.8.1 Energy Distribution

This section gives the results on the energy data for phonemes. The goal of the energy study is to distinguish between vowels, nasals and noisy sounds. This data is used in the next chapter to determine the bit distribution in the coding algorithm.

The phonemic data used in this thesis was obtained from recordings of a number of monosyllabic utterances of a male talker made in an anechoic chamber. These utterances were lowpass filtered to 3600 Hertz. The lowpassed filtered signal was then digitized at 8000 Hertz using the

program DIGITIZ. The digitized data is stored on the INTERDATA computer system disk in data file BURGE.DAT.

For future use, the sentence data in digitized format [146] is stored on the IBM 370 computer system. The data was lowpass filtered to 4000 Hertz and samples at 8000 Hertz. This data is stored in the files listed in Table IX with a description of the data. Representative sonagrams of Table IX are shown in Appendix D.

TABLE IX

SENTENCE DATA

Sentence Description	File
"The pipe began to rust while new"	OSU.ACT10161.SPEECH1
"Add the sum to the product of these three"	OSU.ACT10161.SPEECH2
"Open the crate but don't break the glass"	OSU.ACT10161.SPEECH3
"Oak is strong and also gives shade"	OSU.ACT10161.SPEECH4
"Thieves who rob friends deserve jail"	OSU.ACT10161.SPEECH5
"Cats and dogs each hate the other"	OSU.ACT10161.SPEECH6

The phonemic utterances used in this thesis are shown in Table X. Table X represents a wide variety of speech sounds. The consonants /b/ and /h/ are used to utter syllables of the form consonant-vowel-consonant (CVC) with the consonant /d/ in the final position for the vowels, such

TABLE X

PHONEMIC DATA

No.	Utterance	No.	Utterance	No.	Utterance	No.	Utterance
1	--	41	/y/	81	/ʃ/	121	/hid/
2	/i/	42	/y/	82	/bit/	122	/hId/
3	/i/	43	/m/	83	/bIt/	123	/hɛd/
4	/I/	44	/m/	84	/bɛt/	124	/hæd/
5	/I/	45	/n/	85	/bæt/	125	/hʌd/
6	/ɛ/	46	/n/	86	/bʌt/	126	/hɔd/
7	/ɛ/	47	/ŋ/	87	/hɔt/	127	/hUd/
8	/æ/	48	/ŋ/	88	/bUt/	128	/hud/
9	/æ/	49	/b/	89	/fut/	129	/hɝd/
10	/ʌ/	50	/b/	90	/but	130	/haId/
11	/ʌ/	51	/d/	91	/bɝd/	131	/hɔId/
12	/a/	52	/d/	92	/aIs/	132	/haUd/
13	/a/	53	/g/	93	/bɔI/	133	/hoUd/
14	/ɔ/	54	/g/	94	/baU/	134	/heId/
15	/ɔ/	55	/p/	95	/boU/	135	/hjud/
16	/U/	56	/p/	96	/beIt/	136	/awa/
17	/U/	57	/t/	97	/ju/	137	/ala/
18	/u/	58	--	98	/wIl/	138	/ara/
19	/u/	59	/t/	99	/lIl/	139	/aya/
20	/ɝ/	60	/k/	100	/rIl/	140	/ama/
21	/ɝ/	61	/k/	101	/yIl/	141	/ana/
22	/aI/	62	/h/	102	/mIl/	142	/seŋ/
23	--	63	/h/	103	/nIl/	143	/aba/
24	/aI/	64	/j/	104	/sɛn/	144	/ada/
25	/ɔI/	65	/j/	105	/bIl/	145	/aga/
26	/ɔI/	66	/tʃ/	106	/dIl/	146	/apa/
27	/aU/	67	/tʃ/	107	/gIl/	147	/ata/
28	/aU/	68	/v/	108	/pIl/	148	/aka/
29	/oU/	69	/v/	109	/tIl/	149	/aha/
30	/oU/	70	/ð/	110	/kIl/	150	/aja/
31	/eI/	71	/ð/	111	/hIl/	151	/atʃa/
32	/eI/	72	/z/	112	/jIl/	152	/ava/
33	/jU/	73	/z/	113	/tʃIl/	153	/aða/
34	/jU/	74	/f/	114	/vIl/	154	/aza/
35	/w/	75	/f/	115	/ðæt/	155	/afa/
36	/w/	76	/θ/	116	/aIl/	156	/aθa/
37	/l/	77	/θ/	117	/fIl/	157	/asa/
38	/l/	78	/s/	118	/bæθ/	158	/aʃa/
39	/r/	79	/s/	119	/sIl/		
40	/r/	80	/ʃ/	120	/ʃIl/		

as /hid/. The vowel /a/ is used to utter nonsense syllables of the form vowel-consonant-vowel (VCV) in both initial and final positions, such as /aba/. A set of minimal units using the final form -/Il/ (-ill) is used for the consonants also. Some of the other syllables used are English words. The basic sounds are found in Table II.

The phonemes are analyzed by the algorithms in Appendix B. The energy data is shown in Table XI, normalized by the sound /æ/ for the first 81 phonemes in Table X. The energy in the phoneme /æ/ corresponds to the largest compared to each of the other phonemes. The data is calculated by the program ENERGY. From Table XI, it can be seen that the energy of the prediction residual divides the phonemes into classes by phonemic aggregations.

It is well known that with simple LPC methods [60], the excitation function is a set of periodic pulses or random noises which can be identified as high or low energy excitation functions. However, by using the energy data in Table XI, the phonemes can be grouped into three classes, namely high energy, low energy and noise groups. The high energy group includes the vowels and diphthongs. The plosive, fricative and unvoiced phonemes make up the noise group. The low energy group is composed of glides and nasals. It follows that an ideal excitation signal for speech would enhance perception by considering a three-tier classification rather than the conventional two-source model. This would include a source for vowels, a source for nasals and glides, and a source for fricatives. This is the result of the phoneme energy study of the prediction residual. A normalized energy distribution by phoneme for each sub-band is shown along with the energy bands in Figure 24.

TABLE XI

ENERGY BY PHONEME FOR PREDICTION RESIDUAL

Phoneme	Total Frequency Band	SB1	SB2	SB3	SB4
/i/	.44	.58	.46	.33	.31
/I/	.75	.51	.46	.75	.45
/ɛ/	.84	.65	.47	1.00	.54
/æ/	1.00	1.00	1.00	1.00	1.00
/ʌ/	.72	.67	.57	.45	.41
/a/	.72	.64	.69	.59	.35
/ɔ/	.83	.60	.68	.70	.42
/U/	.24	.23	.25	.20	.15
/u/	.19	.29	.10	.11	.20
/ɚ/	.61	.62	.22	.64	.15
/aI/	1.00	1.00	.79	.68	.65
/ɔI/	.44	.75	.45	.20	.32
/aU/	1.00	1.00	.95	.90	.78
/oU/	.56	1.00	.31	.21	.53
/eI/	.86	1.00	.64	.66	.49
/jU/	.32	.67	.21	.22	.12
/w/	.24	.35	.24	.10	.23
/l/	.24	.29	.08	.10	.29
/r/	.14	.24	.12	.09	.07
/y/	.11	.20	.08	.08	.07
/m/	.34	.65	.25	.22	.19
/n/	.22	.45	.17	.18	.13
/ŋ/	.37	.67	.37	.24	.18
/b/	.24	.49	.11	.14	.17
/d/	.32	.63	.27	.14	.21
/g/	.31	.50	.18	.14	.16
/p/	.18	.27	.08	.07	.08
/t/	.45	.46	.32	.26	.25
/k/	.32	.63	.23	.19	.20
/h/	.45	.46	.24	.31	.31
/j/	.53	.51	.44	.31	.58
/tʃ/	.23	.46	.16	.10	.11
/v/	.16	.29	.13	.09	.09
/ð/	.17	.32	.13	.12	.10
/z/	.24	.44	.17	.19	.15
/f/	.07	.04	.07	.05	.12
/θ/	.11	.21	.09	.07	.05
/s/	.08	.05	.06	.06	.13
/ʃ/	.10	.06	.07	.07	.18

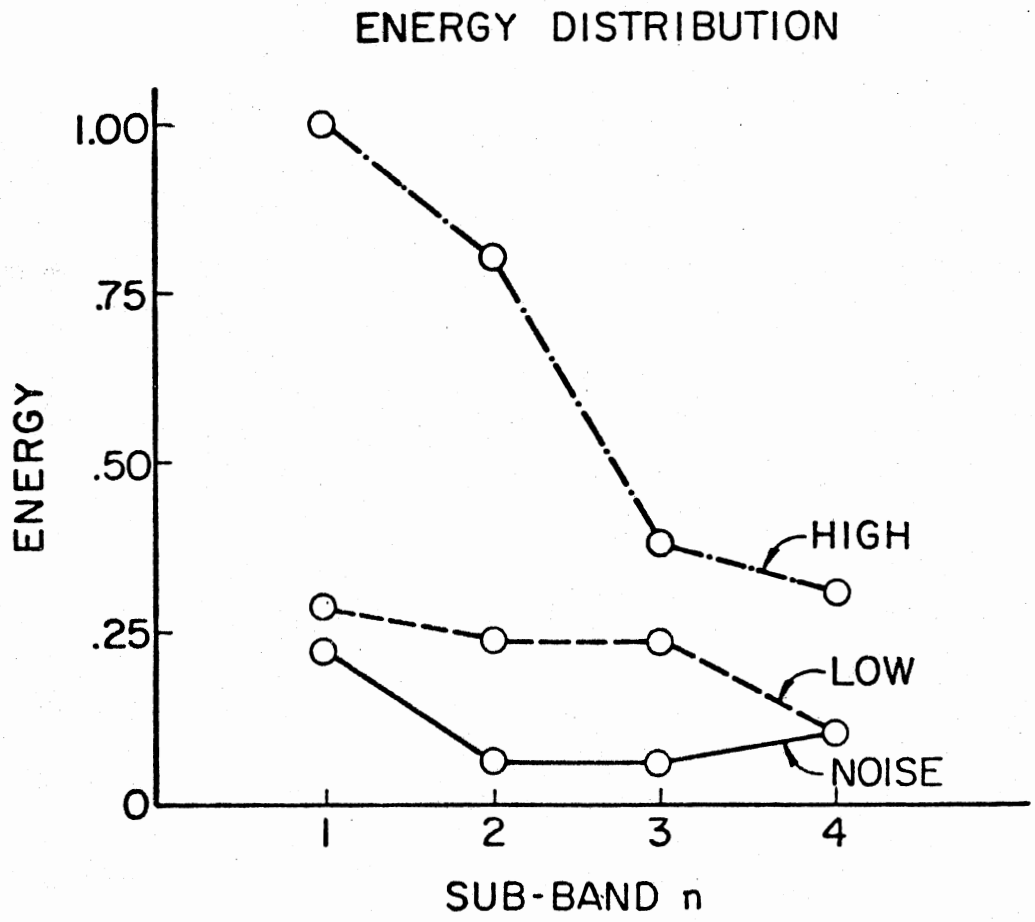


Figure 24. Normalized Energy Distribution by Sub-Band

Based on the above discussion, phonemes can be classified into three energy groups: (1) high energy (HE), (2) low energy (LE) and (3) noise (N). To do this, the normalized residual phoneme energies (second column in Table XI) are the first tabulated; from this, there are clear breaks in the energy levels and therefore three energy groups formed. These breaks are used to identify the threshold values for a particular energy group. For the high energy group, let T_{11} be the threshold value. That is, any phoneme that has normalized residual energy greater than T_{11} is classified into the high energy group. Similarly, T_{22} and T_{33} are the established threshold values for low energy and noise phonemes respectively. The three groupings are given in Table XII. The threshold values T_{ii}, i = 1, 2, 3, can be identified from Figure 24. These are for the entire frequency range.

TABLE XII

PHONEME ENERGY GROUPINGS

Energy Groups	Phonemes
HE	i, I, ɛ, æ, a, ʌ, ɔ, U, u, ʒ
LE	m, n, ŋ, z, w, l, r, y
N	ʃ, f, b, d, g, p, t, k

For the sub-band coding, threshold values need to be computed for each band. Also, each energy group has to be divided into four subgroups corresponding to the four sub-bands. Let E_{in} be the normalized signal energy in the nth frequency band corresponding to the phoneme that is in the ith energy group. This is explicitly shown in Table XIII. For example, E_{12} represents the energy in the second frequency band corresponding to the high energy phoneme (first energy group).

The threshold values for E_{in} (referred hereafter as E_{in}^{T}) in Table XIII will now be established using columns 3, 4, 5 and 6 in Table XI.

TABLE XIII

SYMBOLIC REPRESENTATION OF ENERGY DISTRIBUTION

		Frequency Band			
		1	2	3	4
Energy	H	E_{11}	E_{12}	E_{13}	E_{14}
	L	E_{21}	E_{22}	E_{23}	E_{24}
	N	E_{31}	E_{32}	E_{33}	E_{34}

E_{in} is listed for various phonemes in columns 3, 4, 5 and 6 in Table XI. To make the classification speaker independent, the E_{in} has to be normalized by E_T given in (3.24). Let

$$E_{in}^{T} = \frac{E_{in}}{E_T} \qquad \begin{matrix} i = 1, 2, 3 \\ n = 1, 2, 3, 4 \end{matrix} \qquad (3.25)$$

From this, it is clear that

$$E_{in}^{T} \leq 1.0 \qquad \begin{matrix} i = 1, 2, 3 \\ n = 1, 2, 3, 4 \end{matrix} \qquad (3.26)$$

As before, E_{in}^{T} in (3.25) are tabulated for i = 1, 2, 3 and n = 1, 2, 3, 4. The breaks are established from this tabulation and the threshold values are obtained from these breaks. These are tabulated in Table XIV. The array in Table XIV will be referred hereafter as energy threshold matrix. This matrix will be used in computing the bit allocation scheme, which is discussed in the next chapter.

TABLE XIV

ENERGY THRESHOLD MATRIX

	Frequency Band			
	1	2	3	4
H	.58	.27	1.0	.75
L	.50	.19	1.0	.86
N	.46	.27	1.0	1.00

3.9 Summary

In this chapter, the basis of coding the prediction residual at the rate of 9600 bits/second using the techniques of sub-band coding was presented. Transform coding and sub-band coding wwere discussed along with their relationship. The method of achieving maximum intelligibility based on the Articulation Index was presented. Transitional information of speech along with the relation of speech perception to intelligibility was discussed. Phonemes have been divided into three energy groups so that these can be used in the bit allocation scheme to be discussed in Chapter IV.

CHAPTER IV

ENERGY BASED SUB-BAND CODING ALGORITHM

4.1 Introduction

In this chapter the sub-band coding algorithm, introduced in Chapter III, is examined with the prediction residual as the input source signal. The coding algorithm combines spectral analysis and waveform coding techniques. The combination is intended to provide perceptual enhancement of the speech. The perceptual aspects of speech are a key factor in the bit distribution of the coding algorithm. The bit allocation is established by using the energy groups discussed in the last chapter. For each frame and for each sub-band, the energy $E_n = \frac{1}{N} \sum_k |E_{fn}(k)|^2$ is computed, where E_n indicates the energy corresponding to the nth sub-band in a given frame.

It is well known that most of the spectral density for vocalic sounds and the fundamental frequency are basically found in the sub-band number one (lowest frequency band). The intensity of the energy is substantially high. Spectrogram data can show this. The second formant resides predominantly within the second and third sub-bands and is of the low energy type. These formants determine the transitional cues for certain perceptual effects. The energy of noisy speech sounds, i.e., voiceless fricatives, plosives, etc., has a basic flat spectrum and most of the energy is above 2 kHz. The perceptual effects are discerned in this frequency range. The spectrograms show the intensity of the signal

energy represented by varying shades of gray or black areas [2]. The higher the energy, the darker the area. Spectrograms are included in Appendix D. These figures are included to show the different energy levels associated with different phonemes. From these spectrograms, it can be seen that vowels are typified by dark areas; whereas fricatives, plosives, etc., are shown in a gray area. Although all voiced sounds show a dark color on the spectrogram, Makhoul and Wolf [90] have shown that nasals and glides have a lighter shade when compared to other voiced sounds.

In this study, the energy in each frame of the prediction residual is calculated for each type of phoneme. The bits per sample in each band is allocated on an adaptive basis, using the perceptual criteria discussed in the last chapter. The next section deals with the bit allocation scheme.

The bit allocation method is incorporated into the sub-band coder, which is discussed in Section 4.3. The adaptive strategy is combined with a uniform quantizer with results presented in Sections 4.4 and 4.5. Section 4.6 gives the details of the modules for computational aspects of the coding of the prediction residual.

4.2 Bit Allocation

In this section, the bit allocation scheme is discussed using the energy groupings in Tables XII in Chapter III. In symbolic form, the bit distribution is shown in Table XV for a three-energy level--four sub-band coder, where the rows correspond to the energy levels and the columns correspond to a particular frequency band. For example, k_{23} bits per

sample are assigned for the second energy (LE) band and the third frequency band.

TABLE XV

SYMBOLIC REPRESENTATION OF BIT DISTRIBUTION

	Frequency Band			
	1	2	3	4
High Energy (H)	k_{11}	k_{12}	k_{13}	k_{14}
Low Energy (L)	k_{21}	k_{22}	k_{23}	k_{24}
Noise (N)	k_{31}	k_{32}	k_{33}	k_{34}

The bits are allocated by the empirical formula

$$k_{ij} = \log_2\left(1 + \frac{E_{ij}}{\sigma_j}\right) \qquad \begin{matrix} i = 1, 2, 3 \\ j = 1, 2, 3, 4 \end{matrix} \qquad (4.1)$$

where E_{ij} is the energy from Table XIII and σ_j is a normalization factor determined from the constraint

$$\sum_{j=1}^{4} k_{ij} N_j = C \qquad i = 1, 2, 3 \qquad (4.2)$$

with N_j, $j = 1, 2, 3, 4$, being the number of samples in each band after decimation. The value of C is equal to the total number of bits/frame

minus the number of sync bits per frame. Combining (4.1) and (4.2), it follows that

$$\sum_{j=1}^{4} N_j[\log_2(1 + \frac{E_{ij}}{\sigma_j})] = C \qquad i = 1, 2, 3 \qquad (4.3)$$

where the normalization factor, σ_j, can be determined from (4.3). Equations (4.1), (4.2) and (4.3) define the algorithm.

The normalization factor is included to take into consideration the perceptual aspects of the signal. It is used as a weighting factor for transitional cueing. It has been shown that pitch, formant areas, nasality and affrication are important for speech perception. Within the speech spectrum, these characteristics occur in certain frequency ranges. The power density of speech can indicate this conception, and is discussed below.

The speech power density spectrum is shown in Figure 25. It is clear that most of the energy is below 1000 Hertz. It has been shown by Miller and Nicely [91] that below 1000 Hz, voicing, nasality, and affrication are predominant for determination of the phonemic content. It has been pointed out that given a set of speech signals, a weight factor can be derived when the speech is separated into sub-bands. When these signals are coded properly, there is an advantage of distinguishing certain perceptual effects such as voicing, nasality and affrication. The perceptual effects can be used for calculation of bits for coding.

To compute the normalization factor properly for coding the residual signal, a bit matrix is chosen. The bit distribution that is selected is based on perceptual concepts. This matrix will be referred to as an a priori bit matrix. In addition to perceptual concepts, the a priori bit

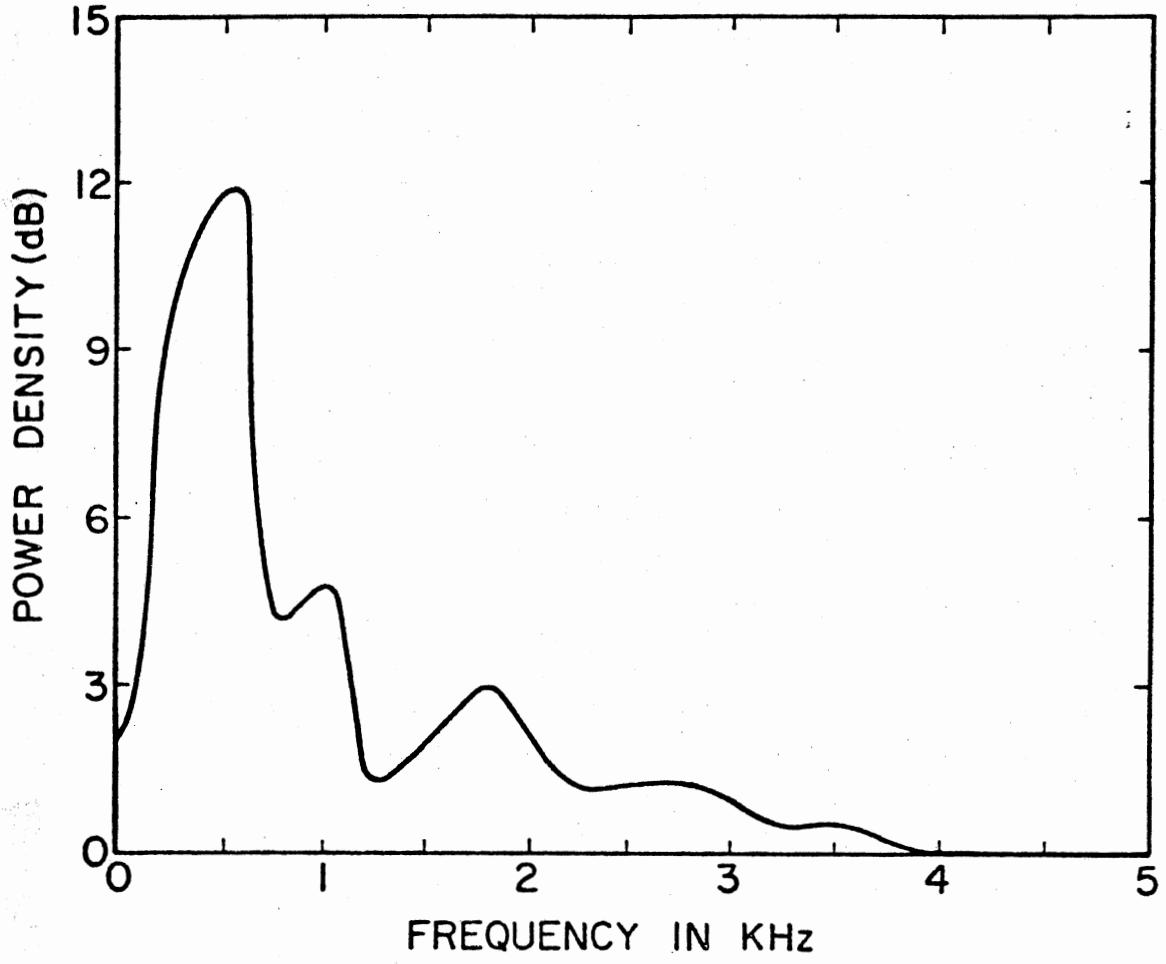


Figure 25. Power Density of Speech Signal (After Noll, 1975)

matrix is selected such that the bit rate is 9600 bits/second for the sub-bands given in Table IV. The matrix is shown in Table XVI, where the entries will be referred to as k_{ij}^{A} to denote the a priori values.

TABLE XVI

A PRIORI BIT MATRIX DISTRIBUTION

		Frequency Band			
		1	2	3	4
High Energy	1	4	3	2	2
Low Energy	2	3	3	3	2
Noise	3	2	3	3	3

The a priori bit matrix is based on experimental results on phonemes. A cursory inspection of Table XVI reveals that the perceptual criteria is preserved. For example, on lower bands where pitch and formant data must be preserved as accurately as possible, a large number of bits per sample are used for encoding, whereas for upper bands where fricatives and noisy sounds are predominant, fewer bits per sample are used. Note that the same number of bits for each energy group is allocated. Also, the a priori bit values (k_{ij}^{A}) are used to compute the normalization factor in (4.1).

When the energy of the speech sound is determined to be high enough, the energy threshold introduced in Chapter III selects the energy matrix (from Table XIII) and a priori bit values (from Table XVI). These are used to calculate the normalization factor from (4.1), and

$$\sigma_j = \frac{E_{ij}^T}{\left(2^{k_{ij}^A}\right) - 1} \qquad \begin{matrix} i = 1, 2, 3 \\ j = 1, 2, 3, 4 \end{matrix} \tag{4.4}$$

where E_{ij}^T is the energy obtained from threshold matrix and k_{ij}^A is obtained from the a priori bit matrix. Figure 26 gives the distribution of $(1/\sigma_j)$ based upon (4.4).

Equation (4.1) can now be used to allocate the bits. It should be pointed out that in using this equation, actual energy values of the signal will be used rather than the threshold values. The following steps are performed to allocate the bits.

1. Spectral estimates are computed for each sub-band.
2. The total energy in the frame for the entire frequency band is computed.
3. E_{ij}'s are computed.
4. Normalization factor, σ_j, is computed
5. The bits are allocated by

$$k_{ij} = \log_2\left(1 + \frac{E_{ij}}{\sigma_j}\right) \qquad \begin{matrix} i = 1, 2, 3 \\ j = 1, 2, 3, 4 \end{matrix} \tag{4.5}$$

where E_{ij} is the energy in the jth sub-band corresponding to the ith energy group and σ_j is the normalization factor from (4.4). Figure 27 gives the flow chart for the bit allocation scheme.

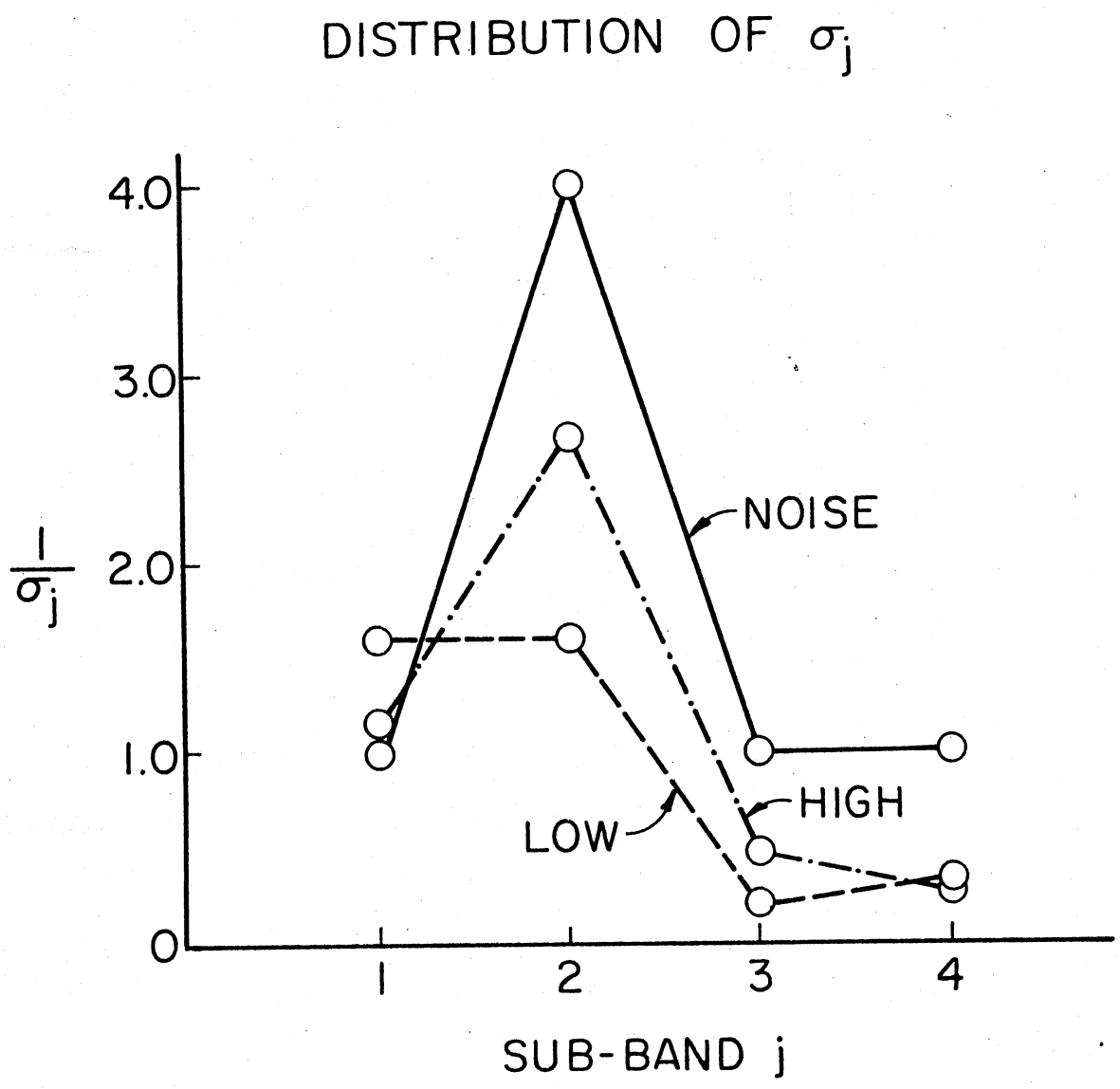


Figure 26. Distribution of Normalization Factor

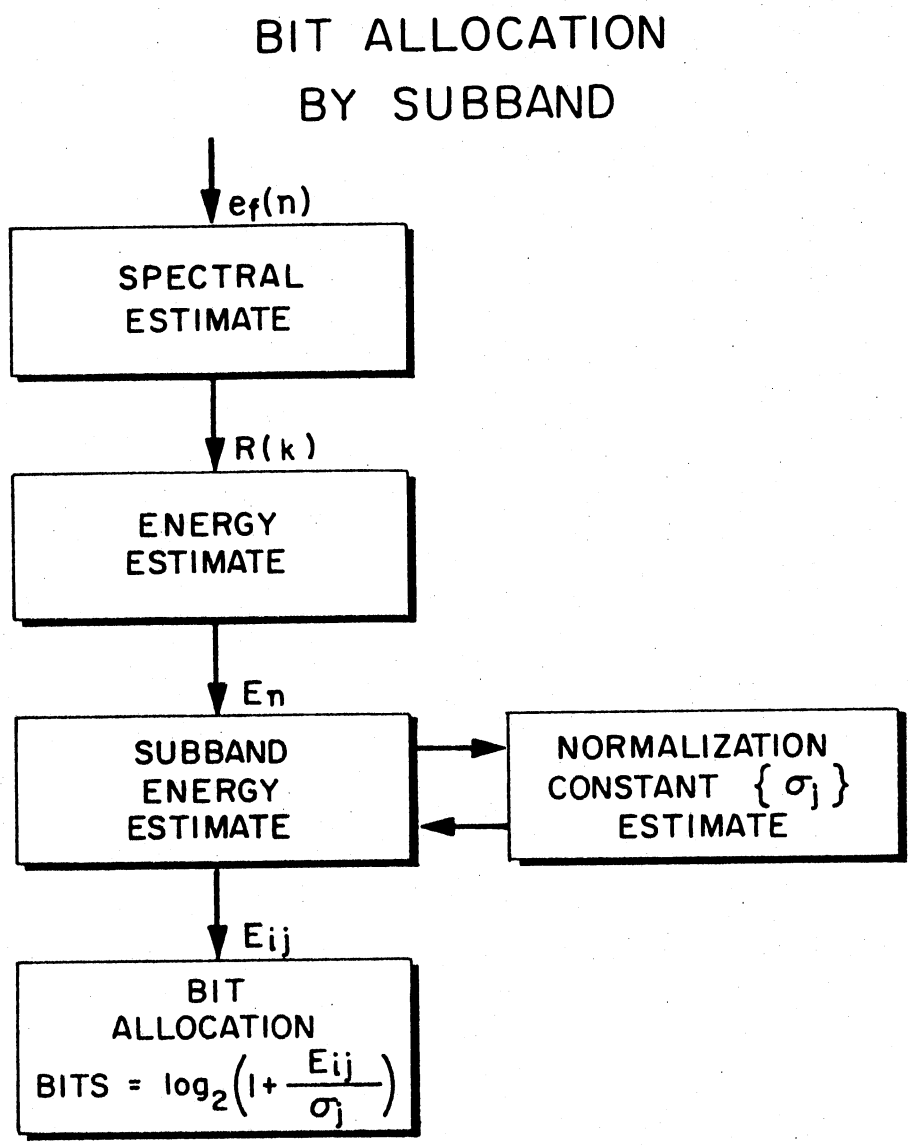


Figure 27. Flow Chart for Bit Allocation

Equation (4.5) has been simulated using the phonemes in Table XII. The bits are averaged for each energy group. The results of the simulations are shown in Figure 28 for each of the three energy groups. Distinctly shown is a separation of the energy groups. Note that the low energy group which contains the nasalic and glide sounds is shown to separate the high energy and noise groups. This separation supports the three-source theory of the residual signal.

Earlier, it was shown that the residual signal parallels glottal excitation. The use of the residual signal for encoding the speech and later exciting the speech synthesizer has several benefits. The bits are minimized in the first and second sub-bands, reducing the necessary transmission rate for these sub-bands. It is unnecessary to transmit twice as many bits for sounds with nasalic, glide or liquid characteristics. On the other hand, the discrimination from the noise is shown to be distinct. The benefit remains clear further, because perceptual criteria will be enhanced in all sub-bands. Discrimination of sounds can be benefited with a minimum bit allocation.

The bit distribution is shown by frame for each phoneme in Figure 29. Again, it is shown that the perceptual criteria is preserved in that the pitch and formant prediminant phonemes receiving a substantial bit allocation and fewer bits are allocated for fricative and plosive phonemes. Noting that the total number of allowed bits per frame is constant, the difference in bits per energy group is adjusted in the synthesis bits. This is discussed in detail in the next section.

4.3 Sub-Band Encoding of the Prediction Residual

The bit allocation scheme was used in the perceptual aspects of

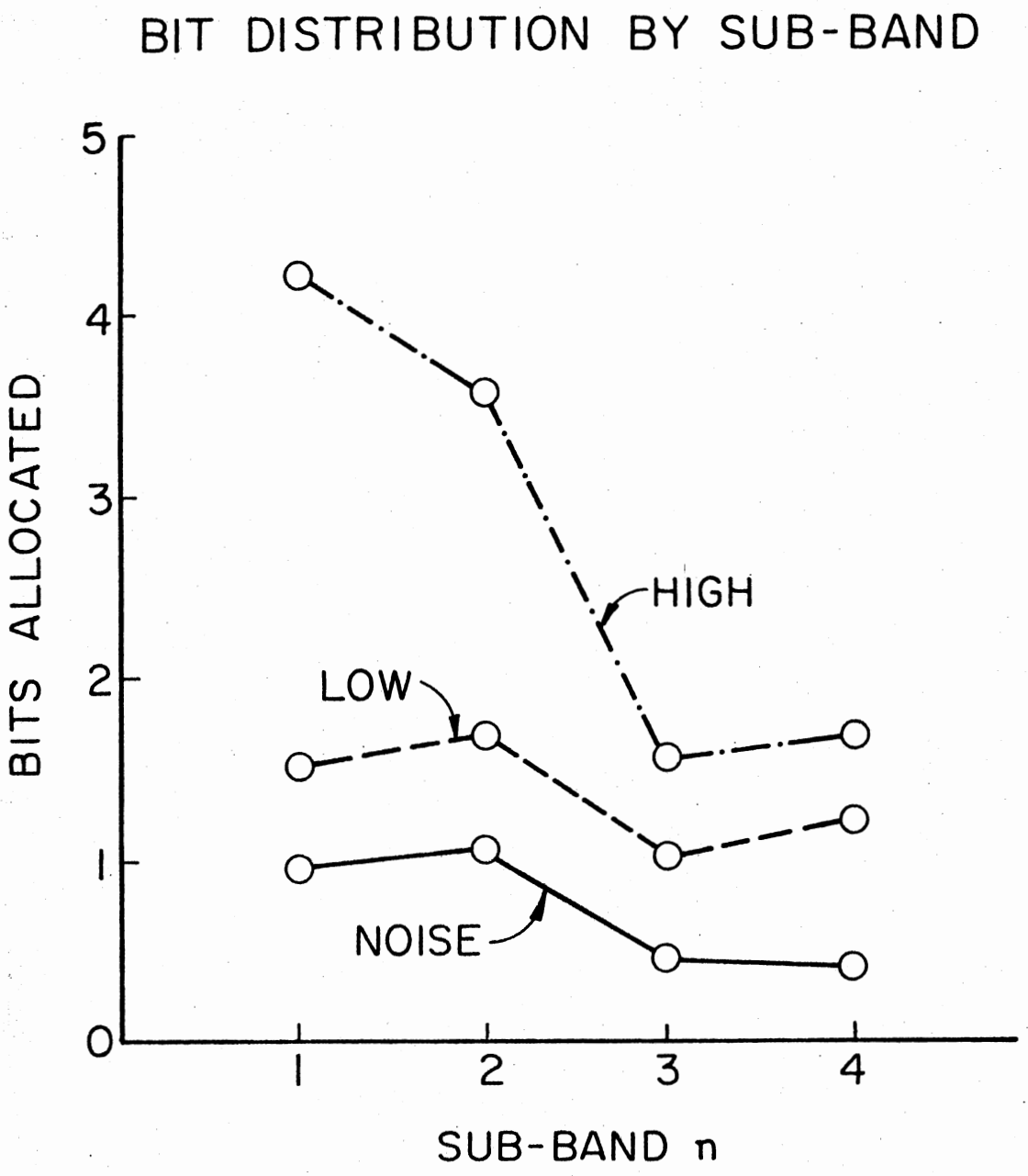


Figure 28. Bit Distribution for Sub-Bands by Energy Bands

speech in sub-band coding of the prediction residual. The sub-band coder partitions the frequency band of the residual signal into four sub-bands by using the bandpass filters. The partitioning of the frequency bands is shown in Figure 20. Each sub-band is low-pass translated, decimated [by the Nyquist interval obtained from (3.15)], and encoded according to the bit allocation scheme discussed above. It has been shown that separate coding of each sub-band accomplished the preferenctial perception criteria for that band [37]. The decoding of each sub-band involves an interpolation and translation back to the original band. The bands are summed to arrive at an estimate of the original residual signal (see Figure 21). This section describes the sub-band coding parameters, the relation of the sub-bands to the Articulation Index and other perceptual criteria discussed in this thesis.

The cutoff frequencies for the sub-band coder are shown in Table XVII. The guideline established for selection of cutoff frequencies is to represent an approximately equal contribution to the Articulation Index. The bands shown in Table XVII represent enough of the important frequencies such that intelligibility is preserved.

TABLE XVII

SUB-BAND CODER CUTOFF FREQUENCIES

Band	Cutoff Frequency (Hz)
1	250 - 500
2	500 - 1000
3	1000 - 1700
4	2000 - 3000

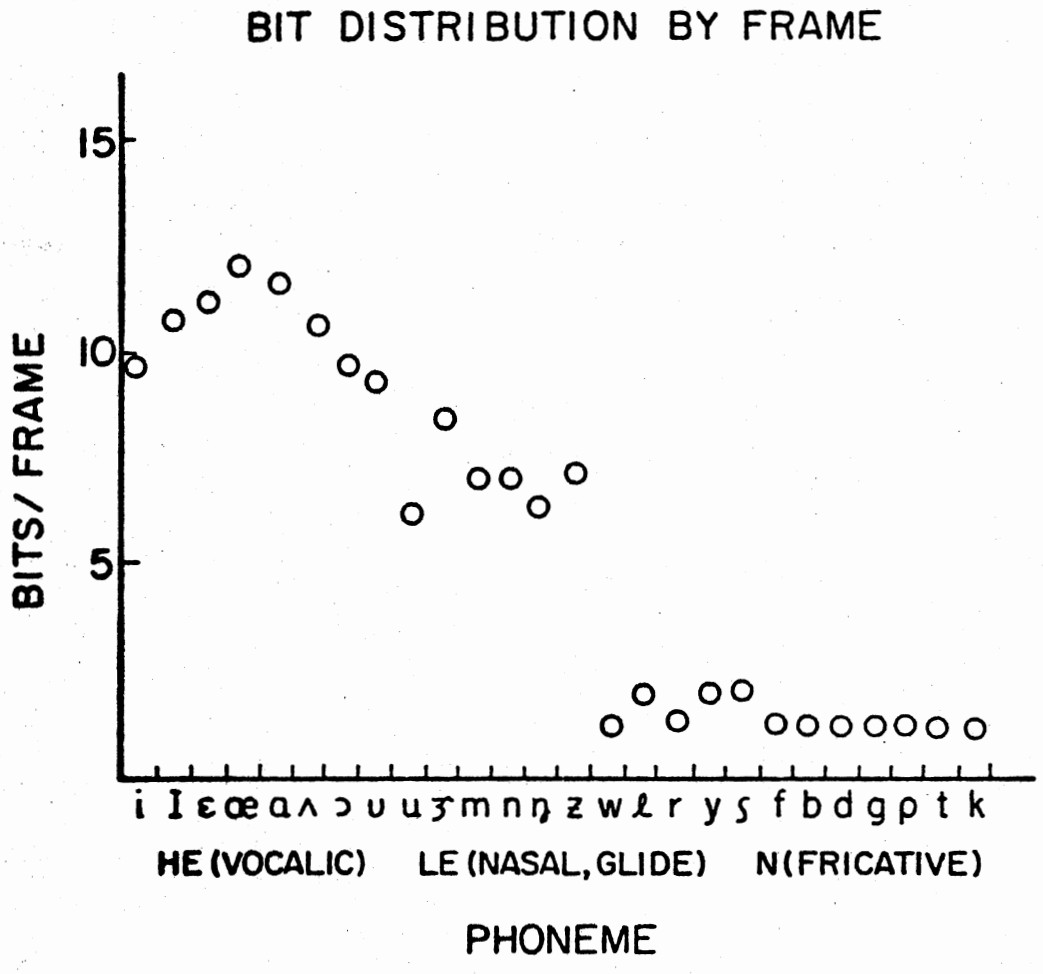


Figure 29. Bit Distribution for Phonemes by Frame

The integer-band sampling scheme [37] was also analyzed at the sampling rate of 8000 Hertz. The technique requires the ratio of upper to lower cutoff frequencies of the sub-band be $(m_i + 1)/m_i$, where m_i is an integer. These bands are related at the bit rate such that the data can be synchronized when multiplexed. Table XVIII is helpful in constructing the sub-bands. Previous authors have given the choice of bands that relate at other sampling rates [36] [37]. Shown in Table XVIII are integer-band sampling cutoff frequencies for an 8000 Hertz sampling rate. The integer decimation ratio is shown in Column 1 for 8000 Hertz. The bandwidths, f_i are indicated in Column 2. The sampling rate, $2f_i$, is shown in Column 3. In Columns 2, 3 and 4, the cutoff frequencies are indicated implicitly. Integer-band sampling is not used in this thesis, and is given here for completeness.

To explain how each band is related, the analysis of the sub-band coder is discussed. The sub-band coder is designed for 9600 bits/second. The transmitted coder parameters include the sub-band coded prediction residual signal, PARCOR coefficients and sync bits. Table XIX represents a breakdown of sub-band coder parameters for the high energy phonemes. Table XX shown sub-band coder parameters relative to the low energy sounds. Table XXI represents those parameters relative to the noise sounds. The difference in Tables XIX, XX, and XXI are the bits allocated and the transmission rates per band, and the sync bits.

It is well known that the decimation rate shown in Column 4 of Tables XIX through XXI represent an integer number of samples available before encoding. These available samples are related to the 9600 bits/second transmission rate. The fractional representation for each frame and sub-band samples are shown in Table XXII.

TABLE XVIII

INTEGER-BAND SAMPLING CUTOFF FREQUENCIES FOR 8000 HERTZ SAMPLING RATE

Decimation Ratio	f_i	$2f_i$	$3f_i$	$4f_i$
1	4000	8000	12000	16000
2	2000	4000	6000	8000
3	1333	2666	3999	5332
4	1000	2000	3000	4000
5	800	1600	2400	3200
6	666	1332	1998	2664
7	571	1142	1713	2284
8	500	1000	1500	2000
9	444	888	1332	1776
10	400	800	1200	1600
11	363	728	1089	1452
12	333	666	999	1332
13	308	616	924	1232
14	286	572	858	1144
15	266	534	798	1064
16	250	500	750	1000
17	235	470	705	940
18	222	444	666	888
19	210	420	630	840
20	200	400	600	800
21	190	380	570	760
22	182	364	546	728
23	174	348	522	696
24	167	334	501	668
25	160	320	480	640
26	154	308	462	616
27	148	296	444	592
28	143	286	429	572
30	133	266	399	532
31	129	258	387	516
32	125	250	375	500

TABLE XIX

SUB-BAND CODER PARAMETERS RELATIVE
TO HIGH ENERGY PHONEMES

Band	Cutoff Frequency (Hz)	Sampling Frequency (Hz)	Decimation Rate	Bits Allocated	Transmission Rate (b/s)
1	250 - 500	500	16	4.0	2000
2	500 -1000	1000	8	3.0	3000
3	1142 - 1700	1142	7	1.5	1700
4	2000 - 3000	2000	4	1.0	2000
Sync and Synthesis					900
					9600 b/s

TABLE XX

SUB-BAND CODER PARAMETERS RELATIVE
TO LOW ENERGY PHONEMES

Band	Cutoff Frequency (Hz)	Sampling Frequency (Hz)	Decimation Rate	Bits Allocated	Transmission Rate (b/s)
1	250 - 500	500	16	2.0	1000
2	500 - 1000	1000	8	2.0	2000
3	1142 - 1700	1142	7	1.0	1142
4	2000 - 3000	2000	4	1.25	2500
Sync and Synthesis					2958
					9600 b/s

TABLE XXI

SUB-BAND CODER PARAMETERS RELATIVE
TO NOISE ENERGY PHONEMES

Band	Cutoff Frequency (Hz)	Sampling Frequency (Hz)	Decimation Rate	Bit Allocated	Transmission Rate (b/s)
1	250 - 500	500	16	1.0	500
2	500 - 1000	1000	8	1.0	1000
3	1142 - 1700	1142	7	.5	571
4	2000 - 3000	2000	4	.5	1000
Sync and Synthesis					6529
					9600 b/s

TABLE XXII

REPRESENTATION OF SAMPLES FOR A FRAME FOR HIGH ENERGY SOUND

Band	Fraction/Frame	Samples/Frame
1	.207	53
2	.312	80
3	.180	45
4	.207	53
Sync and Synthesis	.094	24
	1.000	256 Samples/Frame

The multiplexing (see Figure 21) is simulated on the computer by first appending each of the decimated signals to 256 points per frame by adding zeros. Second, the DFT's of these are taken. Third, the transformed signals are summed. Finally, the IDFT of the summed signal is the multiplexed signal, which has 256 points. The demultiplexing in Figure 21 is simulated using the inverse process. That is, first, the decoded signal is transformed. Second, it is divided into four frequency bands. Third, these frequency coefficients in each band are appended by zeros to get 256 points. Finally, the IDFT of these signals are taken, which gives the demultiplexed signals.

Shown in each of Tables XIX through XXII is a band labeled "Sync and Synthesis." These parameters include synchronization bits and synthesis parameters for the receiver. The synchronization bits include one to

establish the beginning of a frame and three to determine if the frame contains a high, low or noise energy signal. The remaining samples in the sync and synthesis bits are allocated to the PARCOR coefficients for synthesizing the speech.

The PARCOR coefficients are distributed between the range of $|k_j| \leq 1$ and, in most cases, the entire range is not required [28]. It has been shown that the odd-ordered coefficients are somewhat skewed toward the positive side, whereas the even-ordered coefficients are someqhat skewed toward the negative side [28]. The limitation of a quantizer range results in better speech quality for a given number of bits assigned to each coefficient. These parameters have been studied in depth in the literature. Further quantization characteristics of the PARCOR coefficients can be found in [7] [28] [31] [119]. These aspects are used in adjusting the synthesis bits in Tables XIX to XXII, and is outlined below.

Specifically, the following procedure can be used in assigning bits for synthesis parameters. For high energy sounds, 20 bits can be utilized for the 10 PARCOR coefficients. The bit allocation for low energy sounds for the PARCOR coefficients is 70. For the noise energy sounds, the bit allocation is 170. Note that more bits are available for the PARCOR coefficients corresponding to the low energy and the noise signals as compared to the high energy signals. These allocations in synthesis parameters for encoding are adequate. Actual implementation of the bit allocations for the PARCOR coefficients and their effect on the coder has yet to be done.

The find tuning of quantization parameters has yet to be done. The total sub-band system requires many trade-offs in the analysis section.

In the analysis section, allowance must be made for the transmission rate for each sub-band. In the next section, the uniform quantization method is discussed.

4.4 Adaptive Uniform Quantization

The sub-band coder partitions the residual signal into four frequency bands. These banded signals are passed to the quantizer for reduction of information content. The design of the quantizer is determined by the bits allocated as discussed earlier. The amplitude of each residual signal sample is quantized into one of 2^{IBITS} levels, where IBITS is the number of bits allocated for the sub-band. The information content of the digitized signal is IBITS bits per sample. It is shown in Column 6 in Tables XIX through XXI that the information rate for each sub-band is

$$\text{Information Rate} = (\text{Sampling Freq.})_n \times I \text{ bits/second} \qquad I = 1, \ldots, IBITS \qquad (4.6)$$

where $(\text{Sampling Freq.})_n$ is the sampling frequency for the nth sub-band.

After quantization the discrete amplitude level of the signal sample has a value expressed in binary decimal of length IBITS. The value of IBITS ranges from 1 to 5. For example, the value of 2 for IBITS yields amplitude levels of 00, 01, 10 and 11; whereas, a value of 5 would yield 32 five binary length words.

The range of the quantizer is aligned such that the amplitudes of the input residual signal will be within the range of the maximum swing of the output of the quantization levels. The method for accomplishing the assurance that no overload occurs is based on a scheme of analyzing each frame before quantization; i.e., the range of the signal is found before quantization. This is compared to the bits allocated. An adjustment is

made if needed by rounding the bits allocated to the next integer. The method of quantization will be discussed next.

It has been shown that a characteristic of sub-band coded speech is that it has no sample-to-sample correlation [36] [37]. Following this, encoding is best performed by adaptive pulse code modulation (APCM) [109] [121]. Previous encoding based on differential or fixed prediction does not achieve good results for speech using sub-band coders [37]. Each sub-band utilizes a uniform quantizer characteristic. Each sub-band exhibits a different level of energy; therefore, an adaptive uniform quantizer is used utilizing a technique that shrinks and expands the quantizer by sub-band such that the signal is within the range of the maximum quantization level for that sub-band.

To implement the adaptive uniform quantizer, let the step size be denoted by Δ. Figure 30 illustrates the characteristic for the adaptive uniform quantizer [109] and will be idscussed in detail. It is well known that the uniform quantizer level produces error which follows the uniform distribution. That is, the probability density function of the quantization error Q_e is given by

$$f(Q_e) = \frac{1}{\Delta}; \frac{-\Delta}{2} \leq Q_e \leq \frac{\Delta}{2} \tag{4.7}$$

with the variance

$$\sigma^2(Q_e) = \frac{\Delta^2}{12} \tag{4.8}$$

The step size is dependent on the bits allocated.

Let the number of levels be represented by

$$NL = 2^i \qquad i = 1, \ldots, IBITS \tag{4.9}$$

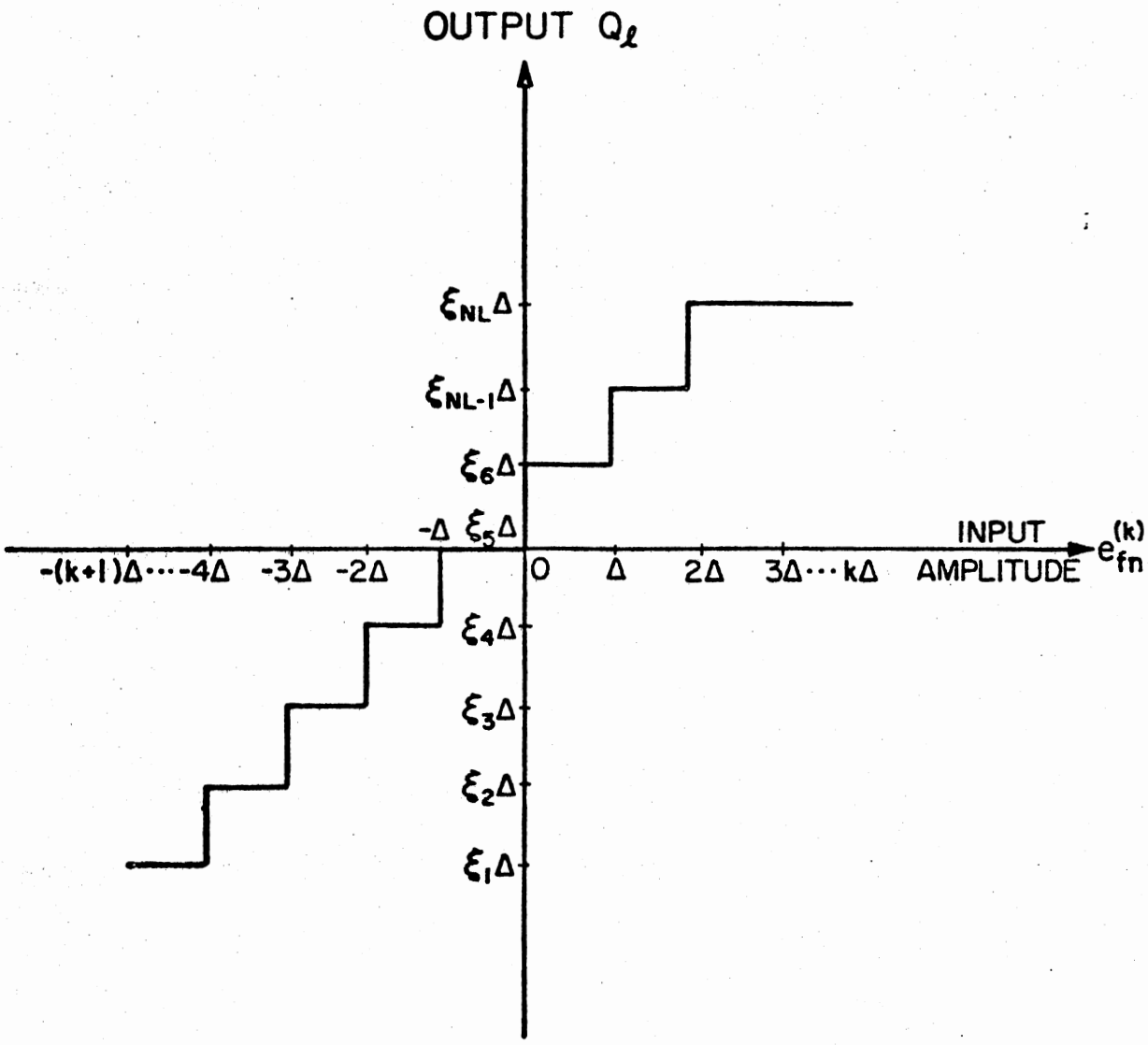


Figure 30. Characteristic of the Adaptive Uniform Quantizer

then

$$\Delta = \frac{2}{NL}[\, e_{fn}(n)\,]_{max} \tag{4.10}$$

where $[\, e_{fn}(n)\,]_{max}$ is the maximum value of the nth sub-band residual signal.

In order to achieve the quantized value, let

$$\underline{\Psi} = [\xi_1, \xi_2, \ldots, \xi_{NL}] \tag{4.11}$$

be an (NL)-vector used to identify the parameters of the quantizer levels such that

$$\underline{Q} = \Delta \cdot \underline{\Psi} \tag{4.12}$$

where the vectors $\underline{Q}$ and $\underline{\Psi}$ are of dimension NL and represent the quantizer values. The entries in $\underline{\Psi}$ are given by

$$\begin{aligned} \xi_\ell &= -\left(\frac{NL}{2} - \ell + 1\right) && 1 \le \ell \le \frac{NL}{2} \\ &= 0 && \ell = \frac{NL}{2} + 1 \\ &= \ell && \frac{NL}{2} + 2 \le \ell \le NL \end{aligned} \tag{4.13}$$

From (4.12) and (4.13), it follows that the quantized level Q_ℓ in $\underline{Q}$ is given by

$$Q_\ell = \Delta \cdot \xi_\ell \tag{4.14}$$

The quantized values of the residual signal are obtained by rounding it to the nearest quantized level, which is used to code the signal.

In the next section, performance measures are discussed for the quantizer and the sub-band coder.

4.5 Signal-to-Noise Ratio Performance Measurements

In the previous section, the quantization is done for the banded prediction residual. In this section, performance measuremetns will be discussed. It has been recognized in the literature that signal-to-noise ratio (SNR) is an inadequate performance measure for speech coding [109]. This idadequacy is realted to the idea that additive white noise is not a good model for error waveforms in speech quantization. Generally, most authors supplement the SNR by subjective and perceptual measurements as a rule.

The SNR is still the single most informative measure for quantizer performance [109]. If the quantizer is designed for maximum SNR, the step size can be chosen according to the probability density function of the signal [122]. However, the SNR improvement is offset by greater idle channel noise for speech [123]. The result is poorer subjective performance [123]. Therefore, to enhance SNR an adaptive quantizing technique is used based on the allocation of bits.

It has been shown that transform coding with adaptive quantizers maximizes SNR and lowers the idle channel noise [81]. Intuitively, sub-band coding should follow under similar conditions. With sub-band coding, the quantization noise of each band is contained within that band and therefore, minimizes the quantization noise of the coded speech [36]. Due to the characteristics of the speech spectrum, the quantization distortion is not equally detectable at all frequencies. This technique offers a means of controlling the quantization noise across the speech spectrum and, therefor a realization of improvement in signal quality [36].

The definition of each objective measure will be discussed next. Perhaps the most common measurement of performance is the conventional (normalized) SNR which is defined as

$$NSNR = -10 \log_{10} \left[\frac{\sum_{k=0}^{N-1} (x(k) - y(k))^2}{\sum_{k=0}^{N-1} x^2(k)} \right] \tag{4.15}$$

where x(k) is the input to the coder and y(k) is the output of the decoder. It is assumed that the numerator represents the noise of the coding technique, such that as the noise decreases a smaller SNR will be the result of the summation in (4.15). The advantage of this quantity is a representation of the normalization of the error between the coder input and the decoder output. For speech there is no perceptual advantage in maximizing the SNR; however, the SNR in (4.15) could be optimized for the autocorrelation of the speech [122].

Another measure similar to (4.15) is the root-mean-square error which is defined as

$$RMSSNR = -20 \log_{10} \left[\sqrt{\frac{\sum_{n=0}^{N-1} (x(n) - y(n))^2}{N}} \right] \tag{4.16}$$

where x(n) and y(n) are defined as before. In (4.16), the error is assumed to be of random nature, and is normalized by the factor N, the number of data points.

A third measure is defined as

$$\text{MSSNR} = -\frac{1}{N}\sum_{n=0}^{N-1} 10 \log\left[\frac{(x(n) - y(n))^2}{x^2(n)}\right] \qquad (4.17)$$

where x(n) and y(n) are expressed as before. The representation in (4.17) defines some measure of error.

The results using (4.15), (4.16) and (4.17) are shown in Table XXIII. These are computed by program SNRCAL (see Appendix B). These results exemplify good coder performance. Note that these simulations are done without bit assignment to PARCOR coefficients. Several phonemes are used in these measurements and they give an adequate measure of the coder. However, the complete simulation should include quantization of all parameters to complete the 9600 bits/second coding algorithm. The next section discusses the computational aspects for coding and decoding the prediction residual.

TABLE XXIII

SIGNAL-TO-NOISE PERFORMANCE MEASUREMENT FOR SEVERAL PHONEMES

Phoneme	RMSSNR	NSNR	MSSNR
/I/	29.2	36.7	18.2
/ɛ/	37.2	36.9	19.1
/æ/	35.1	37.4	17.5
/ʌ/	32.9	34.9	15.2
/a/	30.1	38.4	18.3
/u/	36.8	38.7	17.7
/ɝ/	29.8	38.0	18.4
/aI/	31.3	37.0	18.2
/aU/	34.2	38.4	18.7
/oU/	29.4	37.9	16.8
/eI/	33.4	39.0	17.0

4.6 Computation for Coding the Prediction Residual

The flow chart that gives all the computer modules is given in Figure 31 for coding the residual signal. The data blocks shown in Table XXIV represent data processed and online storage during the computations.

TABLE XXIV

DATA BLOCKS FOR PROCESSING AND STORAGE

Data Block Name	Record Length	Number of Records	Module Used
BURGE.DAT	256	82	DIGITIZ/WINDOW
WINDOW.DAT	256	16	WINDOW/AUTO/LATTIC/INVERS
AUTO.DAT	176	16	AUTO/LATTIC/INVERS
RESIDUAL.DAT	256	16	INVERS/LATTIC/FFTMGR/SUMLPD
SPECTM.DAT	256	16	FFTMGR/RESULT/(PITCH)
BITS.DAT	16	16	FFTMGR/SUMLPD/ENCODE
PHAZ.DAT	256	16	FFTMGR/RESULT
CODE.DAT	256	16	ENCODE/DECODE
SIGNAL.DAT	256	16	DECODE/RESULT
SQNR.DAT	256	16	RESULT
SBAND1.DAT	256	16	SUMLPD/ENCODE
SBAND2.DAT	256	16	SUMLPD/ENCODE
SBAND3.DAT	256	16	SUMLPD/ENCODE
SBAND4.DAT	256	16	SUMLPD/ENCODE

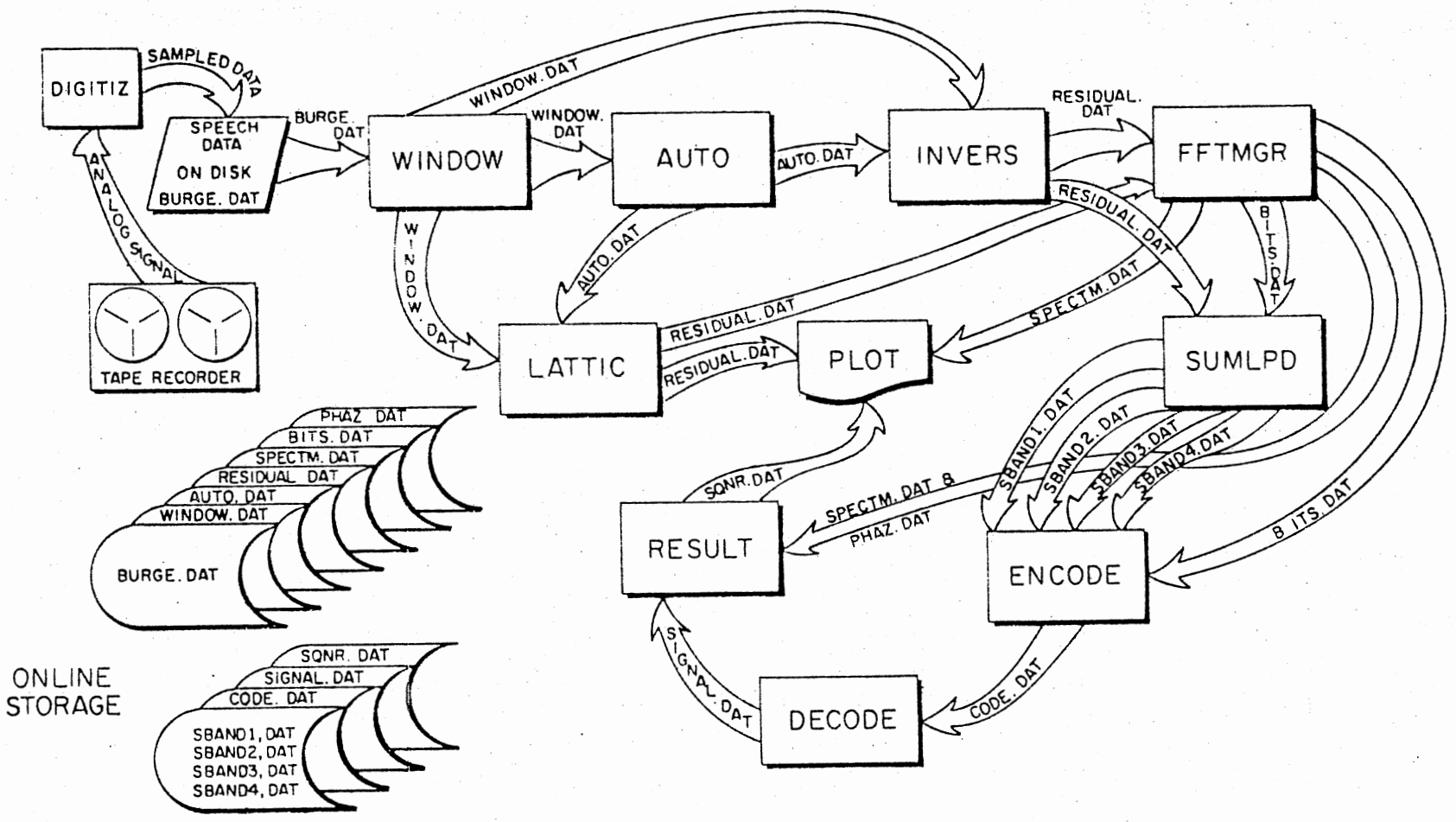


Figure 31. Flow Chart for Coding Residual Signal

The modules are arranged to generate and use the data in Table XXIV on the INTERDATA 70. The tape recorder inputs an analog signal to the computer while DIGITIZ computes a sampled signal and stores the digitized signal on disk in location BURGE.DAT. DIGITIZ is set up to store 4096 points. This program calls an assembly language digitizer and sequence clock for sampling. This program is flexible for sampling any analog signal and storing the signal on disk.

Program WINDOW uses the data on disk, BURGE.DAT. The data is windowed using a 256-point Hamming window. The user has the option of selecting which record of the digitized data to window. The program reports the sequence selected and also scales the data. The window data is written in data block WINDOW.DAT.

Routine AUTO calculates predictor and PARCOR coefficients using Levinson's method [61]. The program uses as input the window data, WINDOW.DAT. The output is an array, AUTO.DAT, containing autocorrelation coefficients, predictor coefficients, cross-correlation coefficients and reflection coefficients.

Routine INVERS uses the data from AUTO, AUTO.DAT, for use in the lattice filter implementation from Equation (2.31) and (2.34). The order of the filter is ten. The output from this program are the residual values. This output is stored in RESIDUAL.DAT. Routine LATTIC is the sames as INVERS except that LATTIC gives the user the option to produce a plot of the speech and prediction residual on CALCOMP.

The FFTMGR module is an FFT manager that includes a bit reversal and unscrambler. The input to this program is the prediction residual, RESIDUAL.DAT. This routine calculates the avergae spectrum, magnitude square and the energy of the prediction residual. It calculates the

energy per sub-band. It uses an a priori estimate of the energy and bits to calculate the normalization factor and bits for each sub-band. The program writes on disk the spectrum, SPECTM.DAT, the bits allocated, BITS.DAT and the phase, PHAZ.DAT. It also gives the user the option for a plot of the spectrum on CALCOMP.

Routine SUMLPD passes the prediction residual through a digital bandpass filter. The signal is modulated, lowpass filtered and decimated as shown in Figure 21. The input to this program is the data file RESIDUAL.DAT. The outputs are the four sub-bands, SBAND1.DAT, SBAND2.DAT, SBAND3.DAT, and SBAND4.DAT.

The signal corresponding to the four sub-bands are encoded using the bits allocated in BITS.DATA by useing the program ENCODE. ENCODE allows for 32 levels of code. In case of non-integer numbers, the quantizer, QUNTIZ, rounds the bits to determine the number of quantizable levels. A uniform quantization is used to determine the code. The output is written in CODE.DAT.

Routine DECODE uses CODE.DAT as input. In the initial frame, the maximum number of quantization levels is determined. This maximum sets the level for the inverse quantizer. Then the signal is decoded and written in file SIGNAL.DAT.

The program RESULT interpolates, modulates and bandpasses the signal, SIGNAL.DAT, for reconstruction. The routine calculates signal-to-noise ratio given by (4.15) and (4.17).

4.7 Summary

In this chapter, the energy based sub-band coding algorithm was presented. The method of allocation of bits was discussed. The design of

the sub-band encoding of the prediction residual was presented. The computational aspects for coding the prediction residual were discussed.

CHAPTER V

SUMMARY AND SUGGESTIONS FOR FURTHER STUDY

5.1 Summary

This thesis investigates an efficient coding of the prediction residual using the technique of sub-band coding at the bit rate of 9600 bits/second. The energy of the prediction residual is used to distribute the bit allocation by sub-bands such that perceptual criteria is preserved. The perceptual criteria is enhanced by transition information embedded in the phoneme connections of speech by a technique that weights the energy based on a normalization factor.

Each sub-band is partitioned such that there is an equitable contribution to the Articulation Index as it is a measure of speech intelligibility. This is discussed in relation to the quality of speech. The perception of speech is described in a qualitative sense. The relationship between the Articulation Index and transitional information is described as a method of discrimination of speech sounds.

The prediction residual is discussed as a parallel to the glottal waveform. The prediction residual is formed by speech through an inverse filter. This is represented as a deconvolution of speech from the vocal tract filter.

The vocal tract filter is modeled as a recursive digital filter using the method of linear prediction. Linear prediction produces the

prediction residual, which is the difference between the actual and predicted speech signals. Because the prediction residual is parallel to glottal excitation, the prediction residual is an ideal pitch extractor.

A novel pitch extraction technique is presented. It is a two-stage method that estimates the residual spectrum and uses time samples of the residual to calculate the approximation of the pitch. The technique calculates a threshold which uses squared samples to extract the pitch within a frame. Also it includes an error check that estimates wide variances of the pitch within each period and is then updated.

The three-tier classification of phonemes is derived from the energy study of the phonemes for the prediction residual. It is shown that the energy of the prediction residual divides the phonemes into classes by phonemic aggregations, namely high energy, low energy and noise groups. The high energy group includes the vowels and diphthongs. The plosive, fricative and unvoiced phonemes compose the noise group. The low energy group is composed of glides and nasals.

The three-tier classification of the energy levels along with the four frequency bands allows for efficient allocation of bits per sample for each band. The above method aids in preserving perceptual criteria and preserves pitch-formant data by the allocation of a large number of bits per sample in the lower bands. Since fricative and noisy sounds are predominant in the upper bands, a smaller number is used in the lower bands. The perceptual criteria is further enhanced by a normalization factor.

The normalization factor is perceptual in nature and is used as a weighting factor for transitional cueing. The derivation of the

normalization factor is discussed. Additional variations are given for the relationship of the three phonemic classes to the normalization factor.

The sub-band coder is designed based on the normalization factor, the energy data, and the bit allocation. The parameters are computed on a frame-by-frame basis. The sub-bands are constructed such that the bit rate of the data from each band can be synchronized when multiplexed at 9600 bits/second. The integer-band sampling scheme is analyzed at the sampling rate of 8000 Hertz for a 9600 bits/second transmission rate. The sub-band coder is designed to transmit the coded prediction residual signal, synthesis parameters and sync bits at the 9600 bits/second rate.

An integral part of the sub-band coder is the quantizer. The encoding of the signal is designed based on adaptive pulse code modulation. Uniform quantization is used. The characteristics of the quantizer are discussed in detail. Performance of the quantizer is described in terms of signal-to-noise ratios (SNR) for objective criterion for quality. The conventional (normalized) SNR is used for representing the error of the coder input and the decoder output. The mean-square SNR is used for an indication of gross error. These SNR measurements are only an indication for quantizer performance. Generally, the SNR must be supplemented by subjective and perceptual measurement as a rule. However, the SNR measurements in this thesis are used without listeners.

In the following, some extensions to the present effort are suggested. Appropriate references are indicated.

5.2 Suggestions for Further Study

5.2.1 PARCOR Coefficient Study of Sensitivity

The PARCOR coefficients introduced in Chapter II have been thoroughly investigated because of their importance to speech analysis and synthesis [9] [28] [31]. The priority is geared toward the synthesis of speech; in that given the prediction residual and PARCOR coefficients, the speech signal can be adequately regenerated. An extension of the present work would enhance present efforts in this area by studying the sensitivity of PARCOR coefficients with respect to the sub-band coding of the prediction residual.

5.2.2 Sub-Band Coding Using Subjective Measurements

The present work can be further advanced by the use of sub-band coding the prediction residual at various bit rates. The synthesized signal would then be used in a comparative study for various bit rates. The perceptual question concerning the method should be geared towards a recording of the synthesized speech so that a set of listeners could hear the results.

5.2.3 Energy Threshold Matrix Study

The introduction of the energy threshold matrix (ETM) in Chapter III requires further study. In this work it is seen that the ETM is highly dependent of perceptual criteria; consequently, several variations would benefit the present work. In some instances, it is necessary to bias the energy group to enhance the perceptual aspects; but this is unknown until

the energy distribution is computed. The results of ETM are dependent on transmission rates; however, given one transmission rate, several ETM may be equally applicable to the coding.

5.2.4 Integer-Band Coding of the Prediction Residual

The integer-band coding method introduced in Chapter IV for use with the prediction residual has not been considered in this thesis. It is simple to implement and would minimize the need for modulators. Previous authors have studied this for speech; however, the subject has not been studied for the prediction residual [36] [37].

5.2.5 Prediction Residual and Noise

A study that would greatly benefit the speech coding area is to mask the prediction residual with white noise. That is,

$$z(k) = e_f(k) + v(k)$$

where $e_f(k)$ represents the discrete samples of the prediction residual signal and $v(k)$ represents the discrete damples of the white noise.

The enhancement of the pitch period markings would be of major importance in this study. Further, the synthesized signal-to-noise ratio performance measurements would also be of interest. The speech waveform has been examined in noise stripping environments; however, the prediction residual in a noise environment has results that are promising [19] [58]. An aid to characterization of the signal would be to use the Laplacian or Gamma distribution, as with the speech. However, these distributions are questionable for the prediction residual since the waveform is different.

Determining the probability distribution of the prediction residual may be a study in itself.

5.2.6 Modeling the Prediction Residual

The prediction residual in this thesis is obtained by inverse filtering the speech signal. Under certain conditions, it is not easy to code the inverse filter; however, if a model was determined that is similar to the signal, it would be of benefit for synthesis. An extension of the work in Chapter II would be to compare the speech and the prediction residual. It would be necessary to identify the essential parameters that can be derived from the residual signal, such as pitch, phase in f_0, formant characteristic and noise between pitch period pulses. The end results would approximate an expression that compares with the actual residual pulse. This in turn could be compared with Flanagan and Rosenberg's work [2] [12] [32].

BIBLIOGRAPHY

(1) Delattre, P. C., A. M. Liberman, and F. S. Cooper. "Acoustic Loci and Transitional Cues for Consonants." The Journal of the Acoustical Society of America, Vol. 27, No. 4 (1955), 769-773.

(2) Flanagan, J. L. Speech Analysis Synthesis and Perception. New York: Springer-Verlag, 1972.

(3) Dudley, H. "Remaking Speech." The Journal of the Acoustical Society of America, Vol. 11 (1939), 165.

(4) Dudley, H., R. R. Riesz, and S. S. A. Watkins. "A Synthetic Speaker." Journal of the Franklin Institute, Vol. 227, No. 6 (1939), 739-764.

(5) Atal, B. S., and S. L. Hanauer. "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave." The Journal of the Acoustical Society of America, Vol. 50, No. 2 (Part 2) (1971), 637-655.

(6) Atal, B. S., and M. R. Schroeder. "Adaptive Predictive Coding of Speech Signals." Bell System Technical Journal (1970), 1973-1990.

(7) Markel, J. D., and A. H. Gray, Jr. Linear Prediction of Speech. New York: Springer-Verlag, 1976.

(8) Markel, J. D. "Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation." IEEE Transactions on Audio and Electroacoustics, Vol. AU-20, No. 2 (1972), 129-137.

(9) Itakura, F., and S. Saito. "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies." Electronics and Communications in Japan, Vol 53-A, No. 1 (1970), 36-43.

(10) Un, C. K., and D. T. Magill. "The Residual-Excited Linear Prediction Vocoder with Transmission Rate Below 9.6 Kbits/s." IEEE Transactions on Communications, Vol. COM-23, No. 12 (1975), 1466-1474.

(11) Markel, J. D., A. H. Gray, Jr., and H. Wakita. Linear Prediction of Speech - Theory and Practice. Santa Barbara, California: Speech Communications Research Laboratory, Inc., SCRL Monograph No. 10, 1973.

(12) Rosenberg, A. E. "Effect of Glottal Pulse Shape on the Quality of Natural Vowels." The Journal of the Acoustical Society of America, Vol. 49, No. 2, Part 2 (1971), 583-590.

(13) Dunn, J. G. "An Experimental 9600-Bits/s Voice Digitizer Employing Adaptive Prediction." IEEE Transactions on Communication Technology, Vol. COM-19, No. 6 (1971), 1021-1032.

(14) Gibson, J. D., S. K. Jones, and J. L. Melsa. "Sequential Adaptive Prediction and Coding of Speech Signals." IEEE Transactions on Communications, Vol. COM-22, No. 11 (1974), 1789-1797.

(15) Cohn, D. L., and J. L. Melsa. "The Residual Encoder - An Improved ADPCM System for Speech Digitization." International Communications Control Conference Record (1975), 30-26 to 30-30.

(16) Goldberg, A. J., A. Arcese, T. McAndres, R. Chueng, and R. Freudbert. Kalman Predictive Encoder. Needham Heights, Massachusetts: GTE Sylvania, Report to Defense Communications Engineering Center, Contract No. DCA 100-74-C-0058, 1975.

(17) McDonald, R. A. "Signal-to-Noise and Idle Channel Performance of Differential Pulse Code Modulation Systems - Particular Applications to Voice Signals." Bell System Technical Journal (1966), 1123-1150.

(18) Qureshi, S. U. H., and G. D. Forney. "Adaptive Residual Coder - An Experimental 9.6/16 KB/S Speech Digitizer." EASCON 1975 Record (1975), 29A-29E.

(19) Berouti, M., and J. Makoul. "High Quality Adaptive Predictive Coding of Speech." International Conference on Acoustic, Speech and Signal Processing Record (1978), 303-306.

(20) Esteban, D., C. Galand, D. Manduit, and J. Menez. "9.6/7.2 KBPS Voice Excited Predictive Coder (VEPC)." International Conference on Acoustics, Speech and Signal Processing Record (1978), 307-311.

(21) Melsa, J. L., D. L. Cohn, J. D. Gibson, R. Kolstad, D. Kopetzky, G. Lauer, and J. Tomkic. Study of Sequential Estimation Method for Speech Digitization. Notre Dame, Indiana: University of Notre Dame, Report to Defense Communications Engineering Center, Contract No. DCA 100-74-C-0037, June 16, 1975.

(22) Atal, B. S., and M. R. Schroeder. "Predictive Coding of Speech Signals and Subjective Error Criteria." IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-27, No. 3 (1979), 247-254.

(23) Cohn, D. L., and J. L. Melsa. "A New Configuration for Speech Digitization at 9600 Bits Per Second." International Conference on Acoustic, Speech and Signal Processing Record (1979), 550-553.

(24) Chang, C. S. "An Improved Residual Encoder for Speech Compression." International Conference on Acoustic, Speech and Signal Processing Record (1979), 542-545.

(25) Magill, D. T., C. K. Un, and S. E. Cannon. Speech Digitization Excitation Study. Menlo Park, California: Standford Research Institute, Report to Defense Communication Engineering Center, SRI Project 1526-8.

(26) Dankberg, M. D., and D. Y. Wong. "Development of a 4.8-9.5 Kbps RELP Vocoder." International Conference on Acoustic, Speech and Signal Processing Record (1979), 554-557.

(27) Viswanathan, R., W. Russell, and J. Makhoul. "Voice-Excited LPC Coders for 9.6 KBPS Speech Transmission." International Conference on Acoustics, Speech and Signal Processing Record (1979), 558-561.

(28) Kang, G. S. Application of Linear Prediction Encoding to a Narrowband Voice Digitizer. Washington, D. C.: Naval Research Laboratory, NRL Report 7774, 1974.

(29) Kryter, K. "Methods for the Calculation and Use of the Articulation Index." The Journal of the Acoustical Society of America, Vol. 34, No. 11 (1962), 1689-1697.

(30) Itakura, F., and S. Saito. "Analysis Synthesis Telephone Based on the Maximum Likelihood Method." The Sixth International Congress on Acoustics Record (1968), C17-C20.

(31) Makhoul. J. "Stable and Efficient Lattice Methods for Linear Prediction." IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-25, No. 5 (1977), 423-428.

(32) Flanagan, J. L. "Some Properties of the Glottal Sound Source." Journal of Speech Hearing Research, Vol. 1 (1958), 99-116.

(33) Rabiner, L. R., B. S. Atal, and M. R. Sambur. "LPC Prediction Error - Analysis of its Variation with the Position of the Analysis Frame." IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-25, No. 5 (1977), 434-442.

(34) Goodman, L. M. "Channel Encoders." Proceedings of the IEEE, (1967) 127-128.

(35) White, C. E. "Bits of Voice." Telecommunications, Vol. 12, No. 4 (1978), 46-48.

(36) Crochiere, R. E., S. A. Webber, and J. L. Flanagan. "Digital Coding of Speech in Sub-Bands." Bell System Technical Journal, Vol. 55, No. 8 (1976), 1069-1085.

(37) Crochiere, R. E. "On the Design of Sub-Band Coders for Low-Bit-Rate Speech Communication." Bell System Technical Journal, Vol. 56, No. 5 (1977), 747-770.

(38) Tribolet, J. M., P. Noll, B. J. McDermott, and R. E. Crochiere. "A Study of Complexity and Quality of Speech Waveform Coders." International Conference on Acoustics, Speech and Signal Processing Record (1978), 586-590.

(39) Barabell, A. J., and R. E. Crochiere. "Sub-Band Coder Design Incorporating Quadrature Filters and Pitch Prediction." International Conference on Acoustics, Speech and Signal Processing Record (1979), 530-533.

(40) Crochiere, R. E. "A Novel Approach for Implementing Pitch Prediction in Sub-Band Coding." International Conference on Acoustics, Speech and Signal Processing Record (1979), 526-529.

(41) Mathews, M. J., J. E. Miller, and E. E. David, Jr. "Pitch Synchronous Analysis of Voiced Sounds." The Journal of the Acoustical Society of America, Vol. 33, No. 2 (1961), 179-186.

(42) Pinson, E. N. "Pitch-Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths." The Journal of the Acoustical Society of America, Vol. 25, No. 8 (1963), 1263-1273.

(43) Sondhi, M. M. "New Methods of Pitch Extraction." IEEE Transactions on Audio and Electroacoustics, Vol. AU-16 (1968), 262-266.

(44) Dubnowski, J. J., R. W. Schafer, and L. R. Rabiner. "Real-Time Digital Hardware Pitch Detector." IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-24 (1976), 2-8.

(45) Noll, A. M. "Cepstrum Pitch Determination." The Journal of the Acoustical Society of America, Vol. 41 (1967), 293-309.

(46) Schafer, R. W., and L. R. Rabiner. "System for Automatic Formant Analysis of Voiced Speech." The Journal of the Acoustical Society of America, Vol. 47, No. 2, Part 2 (1970), 634-648.

(47) Markel, J. D. "The SIFT Algorithm for Fundamental Frequency Estimation." IEEE Transactions on Audio and Electroacoustics, Vol. AU-20 (1972), 367-377.

(48) Miller, N. J. "Pitch Detection by Data Reduction." IEEE Transactions on Audio and Electroacoustics, Vol. ASSP-23 (1975), 72-79.

(49) Gold, B. "Computer Program for Pitch Extraction." The Journal of the Acoustical Society of America, Vol. 34, No. 7 (1962), 916-921.

(50) Gold, B., and L. R. Rabiner. "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain." The Journal of the Acoustical Society of America, Vol. 46 (1969), 442-448.

(51) Rabiner, L. R., M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal. "A Comparative Performance Study of Several Pitch Detection Algorithms." IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-24-, No. 5 (1976), 399-418.

(52) Ross, M. J., H. L. Shaffer, A. Cohen, R. Frendberg, and H. J. Manley. "Average Magnitude Difference Function Pitch Extractor." IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-22 (1974), 353-362.

(53) Maksym, J. N. "Real-Time Pitch Extraction by Adaptive Prediction of the Speech Waveform." IEEE Transactions on Audio and Electroacoustics, Vol. AU-21, No. 3 (1973), 149-153.

(54) McGonegal, C. A., L. R. Rabiner, and A. E. Rosenberg. "A Semi-Automatic Pitch Detector (SAPD)." IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-23 (1975), 570-574.

(55) Wise, J. D., J. R. Caprio, and T. N. Parks. "Maximum Likelihood Pitch Estimation." IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-24, No. 5 (1976), 418-423.

(56) Markel, J. D. "Application of a Digital Inverse Filter for Automatic Formant and F_0 Analysis." IEEE Transactions on Audio and Electroacoustics, Vol. AU-21, No. 3 (1973), 154-160.

(57) Itakura, F., and S. Saito. "On the Optimum Quantization of Feature Parameters in the PARCOR Speech Synthesizer." International Conference on Speech Communication and Processing Record (1972), 434-437.

(58) Boll, S. F. "A Priori Digital Speech Analysis." (Unpub. Ph.D. Dissertation, University of Utah. 1973).

(59) Barnwell, T. D., J. E. Brown, A. J. Bush, and C. R. Patisaul. Pitch and Voicing in Speech Digitization. Atlanta, Georgia: Georgia Institute of Technology, Research Report No. 3-21-620-74BU-1, 1974.

(60) Makhoul, J. "Linear Prediction: A Tutorial Review." Proceedings of the IEEE, Vol. 62, No. 4 (1975), 561-580.

(61) Levinson, N. "The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction." Journal of Mathematics Physics, Vol. 25 (1947), 261-278.

(62) Robinson, E. A., and S. Treitel. "Principles of Digital Wiener Filtering," Geophysics Prospectus, Vol. 15 (1967), 311-333.

(63) Shannon, C. E. "A Mathematical Theory of Communication." Bell System Technical Journal, Vol. 27, No. 3 (July 1948), 379-423, and (October 1948), 623-656.

(64) Flanagan, J. L., M. R. Schroeder, B. S. Atal, R. E. Crochiere, N. S. Jayant, and J. M. Tribolet. "Speech Coding." IEEE Transactions on Communications, Vol. COM-27, No. 4 (1979), 710-737.

(65) Tobias, J. V., ed. Foundations of Modern Auditory Theory, Vol. II. New York: Academic Press, 1972.

(66) Dew, D., and P. J. Jensen. Phonetic Processing - The Dynamics of Speech. Ohio: Charles E. Merrill Publishing Company, 1977.

(67) Rabiner, L. R., and R. W. Schafer. Digital Processing of Speech Signals. New Jersey: Prentice-Hall, 1978.

(68) Jakobson, R., C. G. M. Fant, and M. Halle. Preliminaries to Speech Analysis - The Distinctive Feature and Their Correlates. Cambridge: MIT Press, 1969.

(69) Fant, G. Acoustic Theory of Speech Production. The Hague, The Netherlands: Mouton, 1960.

(70) Jayant, N. S., ed. Waveform Quantization and Coding, New York: IEEE Press, 1976.

(71) Grenander, U., and G. Szegö. Toeplitz Forms and Their Applications. Berkeley, California: University of California Press, 1958.

(72) Durbin, J. "Efficient Estimation of Parameters in Moving - Average Models." Biometrika, Vol. 46, Parts 1 and 2 (1959), 306-316.

(73) Durbin, J. "The Fitting of Time-Series Models." Review of Institution of International Statistics, Vol. 28, No. 3 (1960), 233-243.

(74) McGonegal, C. A., L. R. Rabiner, and A. E. Rosenberg. "A Subjective Evaluation of Pitch Detection Methods Using LPC Synthesized Speech." IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-25, No. 3 (1977), 221-229.

(75) Wightman, F. L., and D. M. Green. "The Perception of Pitch." American Scientist, Vol. 62, No. 2 (1974), 208-215.

(76) Allen, J. B. "Short-Term Spectral Analysis and Synthesis and Modification by Discrete Fourier Transform." IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-25, No. 3 (1977), 235-238.

(77) Allen, J. B., and L. R. Rabiner. "A Unified Theory of Short-Time Spectrum Analysis and Synthesis." Proceedings of the IEEE, Vol. 65, No. 11 (1977), 1558-1564.

(78) Dudley, H. "The Vocoder." Bell Labs Record, Vol. 17 (1939), 122-126.

(79) Jayant, N. S. "Waveform-Coding of Speech." Submitted for publication, Journal of the Acoustical Society of India.

(80) Sambur, M. R. "An Efficient Linear Prediction Vocoder." Bell System Technical Journal, Vol. 54, No. 10 (1975), 1693-1723.

(81) Zelinski, R., and P. Noll. "Adaptive Transform Coding of Speech Signals." IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-25, No. 4 (1977), 299-309.

(82) Huang, J. J., and P. M. Schultheiss. "Block Quantization of Correlated Gaussian Random Variables." IEEE Transactions on Communication System (1963), 291-296.

(83) Brown, J. L., Jr. "Mean Square Truncation Error in Series Expansions of Random Functions." Journal of the Society of Industrial and Applied Math, Vol. 8, No. 1 (1960), 28-32.

(84) Ahmed, N., T. Natarajan, and K. Rao. "Discrete Cosine Transform." IEEE Transactions on Computers, Vol. C-23 (1974), 90-93.

(85) Esteban, D., and C. Galand. "Application of Quadrature Mirror Filters to Split Band Voice Coding Schemes." International Conference on Acoustics, Speech and Signal Processing Record (1971), 191-195.

(86) French, N. R., and J. C. Steinberg. "Factors Governing the Intelligibility of Speech Sounds." The Journal of the Acoustical Society of America, Vol. 19, No. 1 (1947), 90-119.

(87) Beranek, Leo L. "The Design of Speech Communication Systems." Proceedings of the IRE, Vol. 35 (1947), 880-890.

(88) Rabiner, L. R. "Synthesis by Rule." (Unpub. Ph.D. Dissertation, Massachusetts Institute of Technology, 1968).

(89) Lieberman, P. Speech Acoustic and Perception. New York: Bobbs-Merrill Company, Inc., 1972.

(90) Makhoul, J., and J. Wolf. Linear Prediction and the Spectral Analysis. Cambridge: BBN Report No. 2304, 1172.

(91) Miller, G. A., and P. E. Nicely. "An Analysis of Perceptual Confusions Among Some English Consonants." The Journal of the Acoustical Society of America, Vol. 27, No. 2 (1955), 338-352.

(92) Magill, D. T., E. J. Craighill, D. W. Ellis, and C. K. Un. Speech Digitization by LPC Estimation Techniques. Menlo Park, California: Standford Research Institute, ARPA, DDC-Ad-A001931 and AD-785-738, 1974.

(93) Allen, J. B., and R. Yarlagadda. "Digital Poisson Summation Formula and an Application." Submitted for publication, IEEE Transaction on Acoustics, Speech and Signal Processing.

(94) Allen, J. B., and L. R. Rabiner. "Unbiased Spectral Estimation and System Identification Using Short-Time Spectral Analysis Methods." Submitted for publication, IEEE Transactions on Acoustics, Speech and Signal Processing.

(95) Markel, J. D. "FFT Pruning." IEEE Transactions on Audio and Electroacoustics, Vol. AU-19, No. 4 (1971), 305-311.

(96) Atal, B. S., and M. R. Schroeder. "Linear Prediction Analysis of Speech Based on a Pole-Zero Representation." The Journal of the Acoustical Society of America, Vol. 64, No. 5 (1978), 1310-1318.

(97) Elias, P. "Predictive Coding." IRE Transactions on Information Theory, Parts 1 and 2 (1955), 16-33.

(98) Esteban, D., and C. Galand. "32 KBPS CCITT Compatible Split Band Coding Scheme." International Conference on Acoustics, Speech and Signal Processing Record (1978), 320-325.

(99) Noll, P. "A Comparative Study of Various Quantization Schemes for Speech Encoding." Bell System Technical Journal, Vol. 54, No. 9 (1975), 1597-1614.

(100) Wiggins, R. H. Formation and Solution of the Linear Equations Used in Linear Predictive Coding. Bedford, Massachusetts: Mitre Corporation, Report Electronic Systems Division, Report Nos. MTR-2835, ESD-TR-74-301, 1974.

(101) Goldberg, A. J., R. L. Freudberg, and R. S. Bheung. Adaptive Multilevel 16 KB/S Speech Coder. Needham Heights, Massachusetts: GTE Sylvania, Report to Defense Communications Engineering Center, Contract No. DCA 100-76-C-002, 1976.

(102) Goldberg, A. J., and H. L. Shaffer. "A Real-Time Adaptive Predictive Coder Using Small Computers." IEEE Transactions on Communications, Vol. COM-23, No. 12 (1975), 1443-1451.

(103) Qureshi, S. U. H., and G. D. Forney. "A 9.6/16 KB/S Speech Digitizer." International Communication and Control Conference Record (1975), 30-31 to 30-36.

(104) Qureshi, S. U. H., and G. D. Forney. Codex Speech Digitizer Advanced Development Model. Newton, Massachusetts: CODEX Corporation, Report to Defense Communications Engineering Center, Contract No. DCA 100-76-C-0026, 1976.

(105) O'Neal, J. B., and R. W. Stroh. "Differential PCM for Speech and Data Signals." IEEE Transactions on Communications, Vol. COM-20, No. 5 (1972), 900-912.

(106) Burge, L. L., and R. Yarlagadda. "An Efficient Coding of the Prediction Residual." International Conference on Acoustics, Speech and Signal Processing Record (1979), 542-545.

(107) Ahmed, N., and K. R. Rao. Orthogonal Transformations for Digital Processing. New York: Springer-Verlag, 1975.

(108) Oppenheim, A. V., and R. W. Schafer. Digital Signal Processing. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1975.

(109) Jayant, N. S. "Digital Coding of Speech Waveforms - PCM, DPCM and DM Quantizers." Proceedings of the IEEE, Vol. 62, No. 5 (1974), 611-632.

(110) Tribolet, J. M., and R. E. Crochiere. "Frequency Domain Coding of Speech." (Unpub. paper).

(111) Peterson, G. E., and H. L. Barney. "Control Methods Used in a Study of the Vowels." The Journal of the Acoustical Society of America, Vol. 24, No. 2 (1952), 175-184.

(112) Winkler, M. R. "High Information Delta Modulation." IEEE International Convention Record, Part 8 (1963), 260-265.

(113) Ross, M., H. L. Shaffer, A. Cohen, R. Freuberg, and H. J. Manley. "Average Magnitude Difference Function Pitch Extractor." IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-22, No. 5 (1974), 353-361.

(114) Newell, A., J. Barnett, and J. W. Forgie. Speech Understanding System. Amsterdam: North-Holland Publishing Co., 1973.

(115) Papoulis, A. Probability, Random Variables and Stochastic Processes. New York: McGraw-Hill Book Co., Inc., 1965.

(116) Oppenheim, A. V., ed. Applications of Digital Signal Processing. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1978.

(117) Kalman, R. E., and B. S. Bucy. "New Results in Linear Filtering and Prediction Theory." Journal of Basic Engineering, Transactions of the American Society of Mechanical Engineers (1961), 95-108.

(118) Noll, A. M. "Clipstrum Pitch Determination." The Journal of the Acoustical Society of America, Vol. 44, No. 6 (1968), 1585-1591.

(119) Itakura, F., and S. Saito. "Digital Filtering Techniques for Speech Analysis and Synthesis." Seventh International Congress on Acoustics Record (1971), 261-264.

(120) Noll, P. "Effects of Channel Errors of the Signal-to-Noise Performance of Speech-Encoding Systems." Bell System Technical Journal, Vol. 54, No. 9 (1975), 1615-1636.

(121) Papoulis, A. The Fourier Integral and Its Application. New York: McGraw-Hill, 1962.

(122) Noll, P. "A Comparative Study of Various Quantization Schemes for Speech Encoding." The Bell System Technical Journal, Vol. 54, No. 9 (1975), 1597-1614.

(123) Stroh, R. W., and M. P. Paez. "A Comparison of Optimum and Logarithmic Quantization for Speech PCM and DPCM Systems." IEEE Transactions on Communications, Vol. COM-21, No. 6 (1973), 752-757.

(124) Paez, M. C., and Glisson, T. H. "Minimum Mean-Squared-Error Quantization in Speech PCM and DPCM Systems." IEEE Transactions on Communications, Vol. COM-20, No. 4 (1972), 225-230.

(125) Crochiere, R. E., L. R. Rabiner, N. S. Jayant, and J. M. Tribolet. "A Study of Objective Measures for Speech Waveform Coders." Proceedings of the Zurick Seminar on Digital Communications (1978), 261-267.

(126) American Standards Association. American Standard for Preferred Frequencies for Acoustical Measurement. New York, 1960.

(127) Strube, H. W. "Determination of the Instant of Glottal Closure from the Speech Wave." The Journal of the Acoustical Society of America, Vol. 56, No. 5 (1974), 1625-1629.

(128) Atal, B. S. "Automatic Speaker Recognition Based on Pitch Contours." The Journal of the Acoustical Society of America, Vol 52 (1972), 1687-1697.

(129) Rosenberg, A. E., and M. R. Sambur. "New Techniques for Automatic Speaker Verification." IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-23 (1975), 169-176.

(130) Levitt, H. "Speech Processing Aids for the Deaf: An Overview." IEEE Transactions on Audio and Electroacoustics, Vol. AU-21 (1973), 269, 273.

(131) Crochiere, R. E. "An Analysis of 16 Kb/s Sub-Band Coder Performance: Dynamic Range, Tandem Connections, and Channel Errors." Bell System Technical Journal, Vol. 57, No. 8 (1978), 1069-1113.

(132) Beek, B., E. P. Neuberg, and D. C. Hadge. "An Assessment of the Technology of Automatic Speech Recognition for Military Applications." IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-25, No. 4 (1977), 310-322.

(133) Tremain, T. E., J. W. Fussell, R. A. Dean, B. M. Abzug, M. D. Cowing, and P. W. Boudra, Jr. "Implementation of Two Real-Time Narrowband Speech Algorithms." International Conference on Acoustics, Speech and Signal Processing Record (1976), 461-472.

(134) Crochiere, R. E. Personal Communication, August 7, 1978.

(135) Jayant, N. S. Personal Communication, August 7, 1978.

(136) Flanagan, J. L. Personal Communication, August 7, 1978.

(137) Esteban, D., and C. Galand. Personal Communication, August 11, 1978, and April 3, 1979.

(138) Markel, J. D. Personal Communication, February 21, 1979.

(139) Allen, J. B. Personal Communication, February 22, 1979.

(140) Sondhi, M. M. Personal Communication, April 23, 1979.

(141) Tremain, T. E. Personal Communication, June 18, 1979.

(142) Kang, G. S. Personal Communication, June 26, 1979.

(143) Alde, R. Personal Communication, May 16, 1979, and June 18, 1979.

(144) Klein, M. Personal Communication, May 16, 1979, and June 18, 1979.

(145) Sonderegger, R. Personal Communication, August 29, 1978.

(146) Belfield, W. Personal Communication, June 12, 1978.

APPENDIXES

APPENDIX A

DEFINITIONS RELATED TO SPEECH SCIENCE

1. Articulation Index - A weighted fraction representing, for a given speech channel and voice condition, the effective proportion of the normal speech signal which is available to a listener for conveying speech intelligibility. It is computed from acoustical measurements or estimates, at the ear of a listener of the speech spectrum and of the effective masking spectrum of any noise that may be present.
2. Allophone - A manifold acoustic variation of a phoneme.
3. Coding - The means by which an analog waveform is discretized then further represented in one of the well-known methods, e.g., Pulse Code Modulation.
4. Cognate - A complimentary pair of fricatives. One is voiced, the other is unvoiced; however, the place of articulation is the same.
5. Consonant - Those speech sounds which are not exclusively voiced and mouth-radiated. There are fricative, stop and nasal consonants.
6. Communication - The means by which any transmission, emmision or reception of signs, usages or intelligence of any nature is conveyed.
7. Excitation Function - The representation of the glottis in the vocal tract by mathematical modeling in the synthesis of voice.
8. Formant - The resonance component of a speech sound. Generally, it is associated with the phonetic quality of a vowel.
9. Fricative - The speech sound produced by a noise excitation of the vocal tracts. This noise is generated by turbulent air flow at some point along the constriction in the vocal tract. If the vocal cords operate with the noise, the fricative will be voiced; otherwise, it is unvoiced.
10. Glide (liquid) - Those sounds characterized by gliding transition of

the vocal tract and influenced by the context in which the sound occurs, commonly referred to as semi-vowels.

11. Glottis - The orifice between the vocal cords.
12. Intelligibility - The perceptual effect of understanding speech sounds.
13. Language - The set of principles mastered by the speaker in which resides at his grasp an infinite set of sentences. It is a system of human communication based on speech sounds used as arbitrary symbols.
14. Nasal - The group of consonants made with complete closure of the mouth making the radiation sounds come from the nostrils.
15. Phoneme - The basic speech sound element used which serves to keep words apart.
16. Plosives (Stops) - Those speech sounds which begin with complete closure of the lips. The lungs build up pressure behind the closure, suddenly release an explosion marking the voice onset time.
17. Pitch - The difference in the relative vibration frequency of the human voice that contributes to the total meaning of speech, the fundamental frequency.
18. Quality - The ability to identify the character of speech sounds.
19. Place of Articulation - The part of the vocal tract where constriction occurs. Three places of articulation are: labial, alveolar, palatal; i.e., front, middle, and back of the mouth.
20. Speech Perception - The ability of humans to discriminate and differentiate speech sound with their over-learned senses.
21. Suprasegmentals - The features of stress, pitch, intonation, melody, etc., that occurs simultaneously with speech sounds in an utterance.

22. Transitional Cues - The loci of frequency determined by the place of articulation connecting phonemes.
23. Unvoiced - Speech sounds that occur without the vocal cord source operating.
24. Vocal Tract - The acoustic tube which is nonuniform in cross-sectional area beginning with the lips and ending with the vocal cords. For the adult male, it averages 17 centimeters in length and varies from 0 to 20 square centimeters in cross-section.
25. Voiced - Speech sounds that are produced by the vibratory action of the vocal cords.
26. Voice Onset Time - The delay from complete closure of a plosive to the beginning of voicing. Generally averages 25-30 milliseconds.
27. Vowel - Speech sounds produced exclusively by the vocal cord, i.e., voiced, excitation of the vocal tract.

APPENDIX B

COMPUTER PROGRAMS FOR CODING THE PREDICTION RESIDUAL

The flow graph in Figure 24 gives all the programs used in this thesis. These programs were coded for the INTERDATA 70. Each of the modules is discussed below.

B.1 DIGITIZ

This program is an implementation of analog-to-digital (A/D) conversion. It actuates the equipment and the A/D converter which is a part of the computer system. The input is from an analog tape recorder. The output corresponds to the quantized signal with amplitudes of ±10 volts peak-to-peak in steps of 20 millivolts. This data is stored on disk in area BURGE.DAT. The samples are grouped in 16 records sequentially with 256 samples per record.

B.2 LOOK

This program operates on any data set. It was developed as an information tool for scanning the data. It has an option to have the output on a CRT or on a line printer.

B.3 WINDOW

This routine uses a 256-point Hamming window, shifts by 64 points, uses a 256-point window, and the process is continued to the end of the file. The input is the sampled speech data, BURGE.DAT. It conveniently informs the user that the sequence is being windowed. The output is scaled, windowed data that is written in file WINDOW.DAT.

B.4 AUTO

This program calculates the inverse filter coefficients, cross-correlation coefficients, partial correlation coefficients, and auto correlation coefficients. The input is the windowed speech data. A sample of each of the coefficients is printed. They are written in the file AUTO.DAT.

B.5 INVERS and LATTIC

These routines are in implementation of the lattice filter. The input is the windowed speech data. The output is the error signal or the prediction residual. This output is written in the file RESIDUAL.DAT on the disk.

Routine LATTIC provides the user the option of a plot of each frame for the input speech and the prediction residual. The user must also enter the two character names of the sound for the frame desired.

B.6 FFTMGR

This program calculates the Fourier spectrum of the speech input. It calculates the energy per frame, splits this into the predetermined sub-bands for sub-band energy, and it computes the normalization factor and the bit allocation. It uses as input the residual signal and outputs the spectrum, phase and bits. These are written in the files SPECTM.DAT, PHAZ.DAT and BITS.DAT, respectively.

B.7 ENCODE

Routine ENCODE codes a signal based on bits allocated. It uses uniform quantization using the adaptive strategy discussed in the main

part of the thesis to determine the number of levels and rounds the individual samples to the nearest level. The input is the number of bits and the sub-band signal. The coded signal is written in file CODE.DAT.

B.8 DECODE

This routine decodes the integer data in the file CODE.DAT. It determines the largest code level and calculates the allocated bits from this level. It also sets a maximum quantization level. This decoded signal is written in the file SIGNAL.DAT.

B.9 SUMLPD

This routine computes the sub-band prediction residuals using the digital bandpass filters, modulator, lowpass filters, and decimator. The inputs are the signal spectrum and phase. The outputs are the decimated sub-bands. These are written respectively in the files SBAND1.DAT, SBAND2.DAT, SBAND3.DAT, and SBAND4.DAT.

B.10 RESULT

This routine uses the signal to compute signal-to-noise (SNR) ratios. It uses as input the decoded signal and the residual signal. The output is a normalized SNR and an average mean squared SNR. The user has the option of producing a plot. If used, one must input the two-character sound names. The data is written in the file SQNR.DAT.

B.11 SNRCAL

The routine calculates from any two 256-point data arrays the SNR. The input is two arrays of length 256 or less number of points. The

program outputs a mean-squared SNR, a root-mean-square (RMS) SNR and a conventional (normalized) SNR.

B.12 PITCH

This routine estimates the fundamental frequency of a speech utterance. The input is the speech array prediction residual signal and the spectrum of the signal. The program outputs the pitch.

DIGITIZ

```
      IMPLICIT INTEGER*2 (I-N)
      INTEGER*2 BUFF(12000),BLANK,NAME
      DATA F,0,BLANK,LF/5,25,'  ',X'FFFF'/
      DATA NAME/'ZZ'/
      WRITE(5,100)
100   FORMAT('ENTER TRIGGER VALUE THRESHOLD FOR BACKUP ')
300   FORMAT('<CR>(NAME) NAME FOR SPECIFIC IDENTIFICATION')
      CALL INPUT(5,F,0)
 1000  WRITE(5,300)
       READ(5,111) NAME
 111   FORMAT(1A2)
      M=IFIX(F+.5)
      M=M*32
      L=IFIX(0+.5)*32
      WRITE(5,1333)
1333  FORMAT('ENTER CR TO START CONVERSION'/,
     1 'OR "FIN" CR TO STOP PROGRAM')
      READ(5,110)I
110   FORMAT(1A1)
      IF(I.NE.BLANK)GO TO 2000
30    CALL AISQU(9,1,1,N,NSTAT)
      IF (N.LT.M) GO TO 30
      CALL DIGITIZ(BUFF(1),BUFF(12000),L,0,IER)
10    IF(IER.EQ.0) GO TO 10
      WRITE(5,200)
200   FORMAT('DONE')
      DO 15 I=1,7900
      IF(BUFF(I).GT.L) GO TO 25
15    CONTINUE
      GO TO 80
25    I=I-50
      IF(I.LE.0)I=1
      BUFF(I)=NAME
      IEND=I+4095
      DO 20 K=I,IEND
20    BUFF(K)=BUFF(K)/32
26    CALL SYSIO(56,1,IST,ISTDEV,BUFF(I),BUFF(IEND),2,0,0)
      IF(ISTDEV.EQ.0)GO TO 1010
      CALL IOERR(ISTDEV)
      PAUSE 1
      GO TO 26
1014   FORMAT(1A2)
1010   WRITE(5,1014) NAME
      DO 1011 J=I,IEND,7
1011   WRITE(2,1012) BUFF(J),BUFF(J+1),BUFF(J+2),BUFF(J+3),BUFF(J+4)
     +,BUFF(J+5),BUFF(J+6)
1012  FORMAT(1215)
      GO TO 1000
80    WRITE(5,120)
120   FORMAT('NO SIGNAL FOUND RETRY')
      GO TO 1000
2000  ENDFILE 1
      STOP
      END
```

DIGITIZ

LOOK

```
      IMPLICIT INTEGER*2 (I-N)
      DIMENSION ARRAY(256)
      INTEGER*2 BUFF(128)
      EQUIVALENCE (ARRAY(1),BUFF(1))
10    WRITE(5,111)
111   FORMAT('INPUT SEQUENCE NUMBER TO BE DUMPED')
      FLAG=0.0
      CALL INPUT(-5,SN,FLAG)
      IF (FLAG.GE.1.0) GO TO 350
      NS=IFIX(SN+00.5)
      IF(NS.LE.0) GO TO 200
      IREC=32*(NS-1)
      IREC=IREC
      IEND=IREC+31
      REWIND 1
20    CALL SYSIO(92,1,IER,ISTD,BUFF(1),BUFF(128),2,IREC,0)
      CALL IOERR(ISTD)
      IF(IREC.EQ.IEND) WRITE(6,141) BUFF(1)
141   FORMAT(1H1,'SEQUENCE IS ',1A2,/,1X)
      WRITE(6,131) BUFF
131   FORMAT(1X,2016)
      IF(IREC.EQ.IEND) GO TO 10
      IREC=IREC+1
      GO TO 20
350   REWIND 1
360   READ(1,END=10) ARRAY
      WRITE(6,380) ARRAY
380   FORMAT(1X,10F12.3)
      GO TO 360
200   STOP
      END
```

```
      C
      C   ROUTINE WINDOWS DATA BY OVERLAP ADD METHOD
      C   WRITTEN BY J. PERRAULT AND L. BURGE
      C
            IMPLICIT INTEGER*2 (I-N)
            DIMENSION W(64),DATA(256)
            INTEGER*2 INDAT(256)
            EQUIVALENCE (DATA(129),INDAT(1))
    5       WRITE(5,111)
  111       FORMAT('DATA SEQUENCE TO WINDOWED ?')
            CALL INPUT(-5,SN)
            NS=IFIX(SN+0.5)
            IREC=32*(NS-1)
            IBEG=IREC
            IEND=IREC+30
            SCL=6.928
            TWOPI=6.283185
      C
      C   READ IN THE DATA
      C
            REWIND 1
            REWIND 2
            DO 8 I=1,64
    8       W(I)=0.0
   15       CALL SYSIO(92,1,IER,ISTD,INDAT(1),INDAT(256),2,IREC,0)
            IF(ISTD.EQ.0) GO TO 16
            CALL IOERR(ISTD)
            PAUSE 1
            GO TO 5
   16       IF(IREC.GT.IBEG) GO TO 100
      C
      C   REPORT THE SEQUENCE WE ARE ON
      C
            INDAT(1)=INDAT(1)/4
            INDAT(1)=INDAT(1)+128
            WRITE(6,112) NS,INDAT(1)
  112       FORMAT(1H1,'SEQUENCE NUMBER ',I3,' IS ',A2)
            INDAT(1)=0
      C
      C   SCALE THE DATA
      C
            DO 18 I=1,256
   18       DATA(I)=FLOAT(INDAT(I))*SCL
      C
      C   WINDOW THE FIRST BUFFER
      C
            DO 20 I=1,256,64
            N=0
            J=I+63
            DO 20 K=I,J
            PN=FLOAT(K-1)/255.0*TWOPI
            N=N+1
            W(N)=W(N)+0.56-0.46*COS(PN)
   20       DATA(K)=DATA(K)*W(N)
            GO TO 300
      C
  100       IF(IREC.EQ.IEND) GO TO 150
            DO 110 I=1,256
  110       DATA(I)=FLOAT(INDAT(I))*SCL
            DO 120 I=1,256,64
            N=0
            J=I+63
            DO 120 K=I,J
            N=N+1
  120       DATA(K)=DATA(K)*W(N)
            GO TO 300
  150       DO 160 I=1,256
  160       DATA(I)=FLOAT(INDAT(I))*SCL
            DO 180 I=1,256,64
            N=0
            N1=65
            J=I+63
            DO 180 K=I,J
            N=N+1
            N1=N1-1
            DATA(K)=DATA(K)*W(N)
            PN=FLOAT(257-K)/255.0*TWOPI
  180       W(N1)=W(N1)-0.54+0.46*COS(PN)
      C
      C   WRITE OUT THE BUFFER
      C
  300       CALL SYSIO(56,2,IER,ISTD,DATA(1),DATA(256),4,0,0)
            IF(ISTD.EQ.0) GO TO 310
            CALL IOERR(ISTD)
            PAUSE 2
            GO TO 5
  310       IF(IREC.EQ.IEND) GO TO 400
            IREC=IREC+2
            GO TO 15
  400       STOP
            END
```

WINDOW

```
      SUBROUTINE AUTO
      IMPLICIT INTEGER*2(I-N)
C       PROGRAM CALCULATES INVERS FILTER COEFFICIENTS BY
C       LEVINSON'S METHOD
C       N - NO. OF POINTS
C       X - VECTOR OF SAMPLED SPEECH PTS
C       M - ORDER OF INVERS FILTER
C      A - VECTOR OF INVERSE FILTER COEFFICIENTS
C      ALPHA - VECTOR OF CROSS CORRELATION COEFFICIENTS
C      RC - VECTOR OF REFLECTION COEFFICIENTS
C      R - AUTOCORRELATION COEFFICIENTS
C      ARRAY - VECTOR TO HOUSE ALL COEFFICIENTS CALCULATED
      DIMENSION  R(11),X(256),A(11),ALPHA(11),RC(11),ARRAY(44)
      EQUIVALENCE (ARRAY(1),R(1)),(ARRAY(12),A(1)),
     +(ARRAY(23),ALPHA(1)),(ARRAY(34),RC(1))
      N=256
      NS=0
C       READ WINDOWED DATA
      REWIND 1
      REWIND 2
5     READ(1,END=99) X
      M=10
11    MP = M + 1
      DO 10 K = 1,MP
      R(K) = 0.0
      L = N - K + 1
      DO 10 NP = 1,L
      NPK =NP+K-1
10    R(K) = R(K)+X(NP)*X(NPK)
      A(1) = 1.
      ALPHA(1) = R(1)
      IF(R(1).EQ.0) GO TO 60
      RC(1) = -R(2)/R(1)
      A(2) = RC(1)
      ALPHA(2) = R(1) + R(2)*RC(1)
      IF(M.EQ.1) GO TO 60
      DO 40 MINC = 2,M
      S = 0.0
      DO 20 IP = 1,MINC
      MNC=MINC-IP+2
20    S = S + R(MNC)*A(IP)
      RC(MINC) = -S/ALPHA(MINC)
      MH = MINC/2 + 1
      DO 30 IP = 2,MH
      IB = MINC - IP + 2
      AT = A(IP) + RC(MINC)*A(IB)
      A(IB) = A(IB) + RC(MINC)*A(IP)
30    A(IP) = AT
      A(MINC+1) = RC(MINC)
      ALPHA(MINC+1) = ALPHA(MINC)-ALPHA(MINC)*RC(MINC)*RC(MINC)
      IF(ALPHA(MINC)) 50,50,40
40    CONTINUE
50    M = MINC - 1
60    CONTINUE
      WRITE(2) ARRAY
      M1 = M + 1
      WRITE(5,101)
101   FORMAT(1H0,'AUTOCORRELATION TERMS')
100   FORMAT(1H0,5E16.5/5E16.6)
      WRITE(5,100) (R(I),I=1,M1)
102   FORMAT(1H0,'INVERSE COEFFICIENTS')
      WRITE(5,102)
      WRITE(5,100) (A(I),I=1,M1)
103   FORMAT(1H0,'CROSS CORRELATION TERMS')
      WRITE(5,103)
      WRITE(5,100) (ALPHA(I),I=1,M1)
104   FORMAT(1H0,'REFLECTION COEFFICIENTS')
      WRITE(5,104)
      WRITE(5,100) (RC(I),I=1,M)
      NS=NS+1
      WRITE(5,119) NS
119   FORMAT(1H0,'PRESENTLY CALCULATING AND PROCESSING
     + CORRELATION DATA FOR SPEECH SECTION',I5)
      GO TO 5
99    STOP
      END
```

AUTO

INVERS

```
      IMPLICIT INTEGER*2(I-N)
C     IMPLEMENTATION OF LATTICE FILTER BY L. L. BURGE
C     A1 - INPUT OF SPEECH SAMPLES
C     N - NO. OF POINTS
C     M - ORDER OF FILTER
C     RC - REFLECTION COEFFICIENTS
C     RO - RESIDUAL OF SPEECH PTS OUTPUT
C       B - VECTOR OF BACKWARD PREDICTED SAMPLES
C       C - TEMPORARY VECTOR FOR SAMPLES
      DIMENSION F(11),A1(256),B(11),ALPHA(11),RC(11),ARRAY(44)
      DIMENSION C(11,256),B(11),SI(11)
      DIMENSION RO(256)
      EQUIVALENCE (ARRAY(1),K(1)),(ARRAY(12),A(1))
      EQUIVALENCE (ARRAY(23),ALPHA(1)),(ARRAY(34),RC(1))
      M=10
      N=256
      NS=0
C      READ WINDOWED DATA
      REWIND 1
      REWIND 2
      REWIND 3
5     READ(1,END=99) A1
      READ(2) ARRAY
      B(1) = 0.0
      DO 40 K = 1,N
      C(1,K)=A1(K)
      IF(K.GT.1) B(1)=A1(K-1)
      DO 30 J=1,M
      SI(1)=A1(K)
      SI(J+1)=SI(J)+RC(J)*B(J)
      C(J+1,K)=B(J)+RC(J)*SI(J)
      IF (K.EQ.1) GO TO 20
      B(J+1)=C(J+1,K-1)
      GO TO 30
20    B(J+1)=0.0
30    RO(K)=SI(J+1)
40    CONTINUE
      WRITE(3) RO
      M1=1+50
      WRITE(5,100) (RO(I),I=1,M1)
100   FORMAT(1H0,'RESIDUAL VALUES FROM LATTICE FILTER'/
     A(10F12.6))
      NS=NS+1
      WRITE(5,500) NS
500   FORMAT(1H0,'PRESENTLY PROCESSING DATA THRU LATTICE FILTER
     +    SPEECH SECTION ',I5)
      GO TO 5
99    STOP
      END
```

INVERS

LOWPAS

```
      SUBROUTINE LOWPAS(DEMOD,IBAND,I,NB,L,FASE)
C
C        COMPUTES IDEALIZED LOWPASS FILTER
C
      IMPLICIT INTEGER*2(I-N)
      INTEGER*2 IBAND(1)
      REAL MAG(256)
      DIMENSION DEMOD(1),Y(256),FASE(1)
      IZERO=256-IBAND(I)
      IPASS=IBAND(I)
      WRITE(5,100)
100   FORMAT(1HJ,'ROUTINE LOWPASS PROCESSING')
      CALL FFFT(DEMOD,Y,NB,L,1)
      DO 20 K=1,256
      MAG(K)=DEMOD(K)*DEMOD(K)+Y(K)*Y(K)
      FASE(K)=ATAN2(Y(K),DEMOD(K))
140   FORMAT(1H ,'REAL  IMAG  MAG  PHASE   = ',4E15.4)
      DEMOD(K)=SQRT(MAG(K))
20    CONTINUE
      WRITE(5,140) (DEMOD(K),Y(K),MAG(K),FASE(K),K=1,256,32)
      DO 30 K=IPASS,IZERO
30    DEMOD(K)=0.0
      RETURN
      END
C
C
```

ENCODE

```
      C         SUBROUTINE ENCODE
      C
      C         ROUTINE CODES A SIGNAL BASED ON ALLOCATED BITS
      C         USES UNIFORM QUANTIZER, WRITTEN BY L. L. RUDGE
      C
            IMPLICIT INTEGER*2(I-N)
            INTEGER*2 KODE(32), IKODE(256)
            DIMENSION DEC(256), BITS(4), QDEC(256)
            DATA KODE/1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,
           +19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32/
            I1=1
            N=256
            DO 10 I=1, N
            DEC(I)=0.0
            QDEC(I)=0.0
      10    IKODE(I)=0
      C
      C         READ DATA
      C
            READ(1) 1
            READ(2) 2
            READ(3) 3
            READ(4) 4
            READ(7) 7
            READ(8) 8
            READ(7) BITS
            WRITE(5,110) (BITS(J), J=1, 4)
      110   FORMAT(1H ,'ALLOCATED BITS'/F5.2,'  FOR SUBBAND ONE'/
           +F5.2,'  FOR SUBBAND TWO'/F5.2,'  FOR SUBBAND THREE'/
           +F5.2,'  FOR SUBBAND FOUR')
            READ(1) DEC
            DO 30 I=1, N
      30    QDEC(I)=DEC(I)
            BIT=BITS(I1)
            WRITE(5,120) (DEC(I), I=1, N)
      120   FORMAT(1H ,'INPUT SIGNAL ', 12F8.2)
      C
      C         FIND MAXIMUM OF INPUT SIGNAL
      C
            CALL MAX(QDEC, N, BIG, NUM)
            I1=I1+1
      C
      C         CALL THE QUANTIZER
      C
            CALL QUNTIZ(DEC, BIT, QDEC, BIG, KODE, IKODE)
            WRITE(8) IKODE
```

ENCODE

```
            WRITE(5,100) (IKODE(K), K=1, N)
      100   FORMAT(1H ,'ENCODED SIGNAL ', 8I4)
            IF(I1.EQ.5) GO TO 99
            GO TO 1
      99    WRITE(5,999)
      999   FORMAT(1H1///'*******  DONE  *******')
            STOP
            END
      C
      C
```

CHEKEN

```
            SUBROUTINE CHEKEN(E, SEN, IN, ESUB)
      C     THIS PROGRAM CHECKS FOR HIGH ENERGY VS LOW ENERGY SOUNDS
      C     WRITTEN BY L. L. RUDGE
      C     IF PREVIOUS ENERGY WAS MINIMAL THE ENERGY IS REJECTED
      C      E - VECTOR OF ENERGY POINTS
      C      SEN - VECTOR OF TOTAL ENERGY FOR N PTS
      C      IN - FRAME NUMBER
      C      ESUB - SUB-BAND ENERGY
            IMPLICIT INTEGER*2(I-N)
            DIMENSION E(256), SEN(1), ESUB(1)
            AVTE=0.0
            DO 10 I=1, IN
      10    AVTE=SEN(I)+AVTE
            AVTE=AVTE/FLOAT(IN)
            AVTE=AVTE+75.0
            IF(AVTE.GT.E(256)) GO TO 30
            E(256)=E(256)/16.0
            DO 20 I=1, 4
      20    ESUB(I)=ESUB(I)/16.0
      30    RETURN
            END
```

QUNTIZ

```
      SUBROUTINE QUNTIZ(QDEC,BITS,QDEC,BIO,KODE,IKODE)
C
C        ROUTINE QUANTIZES AND CODES SIGNAL BY L. L. BURGE
C
      IMPLICIT INTEGER*2(I-N)
      INTEGER*2 IKODE(1),KODE(1)
      DIMENSION QNEW(33),QDEC(1),DEC(1)
      N=256
      DO 5 I=1,33
    5 QNEW(I)=0.0
C
C        ROUND BITS TO INTEGER, DETERMINE NO. LEVELS
C
      ROUND=BITS+1.0
      IBITS=IFIX(ROUND)
      IQNL=2**IBITS
      IQNL1=IQNL+1
      IF(IQNL .GT. 32) IQNL=32
      QNL=FLOAT(IQNL)
C
C        DETERMINE INCREMENT IN EACH LEVEL
C
      QINC=(2.0*BIO)/QNL
      IQNLH=IQNL/2
C
C          SET VALUE FOR LEVELS
C
      DO 10 K=1,IQNLH
      K1=IQNL1-K+1
      K2=IQNL-K+2
      QNEW(K2)=FLOAT(K1)*QINC
      QNEW(K)=-FLOAT(K1)*QINC
   10 CONTINUE
      WRITE(5,110) IBITS,IQNL,QINC,BIO
  110 FORMAT(1H ,'QUANTIZATION DATA '//' BITS FOR CODING
     + AVAILABLE ',I6/' NUMBER OF LEVELS ',I6/'INCREMENT
     + VALUE ',F6.3/' MAX VALUE OF SIGNAL ',F6.3)
      WRITE(5,100) (QNEW(K),K=1,IQNL)
  100 FORMAT(1H ,' QUANTIZED INCREMENTS '/4F10.3)
C
C        ROUND SIGNAL TO NEAREST LEVEL
C
      DO 30 K=1,N
      DO 60 J=1,IQNL
      IF (QNEW(J+1) .LE. DEC(K)) GO TO 60
      CALL CLAMP(QNEW,J,QDEC,K,DEC,IKODE,KODE)
      GO TO 30
   60 CONTINUE
   30 CONTINUE
      RETURN
      END
C
C
C
C
```

QUNTIZ

CLAMP

```
      SUBROUTINE CLAMP(QNEW,J,QDEC,K,DEC,IKODE,KODE)
C
C        CLAMPS SIGNAL TO NEAREST LEVEL BY L. L. BURGE
C
      IMPLICIT INTEGER*2(I-N)
      INTEGER*2 IKODE(1),KODE(1)
      DIMENSION QDEC(1),QNEW(1),DEC(1)
      CHK1=ABS(QNEW(J+1)-DEC(K))
      CHK2=ABS(DEC(K)-QNEW(J))
      IF(CHK1 .LE. CHK2) GO TO 10
      QDEC(K)=QNEW(J)
      IKODE(K)=KODE(J)
      GO TO 20
   10 QDEC(K)=QNEW(J+1)
      IKODE(K)=KODE(J+1)
   20 RETURN
      END
C
C
```

DECODE

```
      C          SUBROUTINE DECODE
      C
      C          ROUTINE DECODES CODED SIGNAL, WRITTEN BY L. L. BURGE
      C
            IMPLICIT INTEGER*2(I-N)
            INTEGER*2 KODE(256), IKODE(32), KODET(256)
            DIMENSION SIGNAL(256)
            DATA IKODE/1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,
           +19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32/
              N=256
              II=1
      C
      C          READ DATA, FIND MAX IN ARRAY
      C
            REWIND 1
            REWIND 2
      1     READ(1, END=99) KODE
            II=II+1
            DO 20 I=1, N
      20    KODET(I)=KODE(I)
            CALL MAX(KODET, N, IBIG, NUM)
            BIG=FLOAT(IBIG)
            WRITE(5, 110) IBIG, NUM, N
            WRITE(5, 120) (KODE(I), I=1, N)
      120   FORMAT(1H ,' CODED SIGNAL  ', 815)
      110   FORMAT(1H ,'VALUES FROM MAX PROGRAM', 316)
      C
      C          SET LEVEL FOR INVERSE QUANTIZER
      C
            DO 10 INC=1, 16
            IOLD=INC*2+1
            IF(IBIG.EQ.IOLD) BITS=3.32193*ALOG10(BIG)
      10    IF(IBIG.EQ.INC*2) BITS=3.32193*ALOG10(BIG)
            IF(BITS.GT.4.8 AND.BITS.LT.5.2) QLEVEL=10.8
            IF(BITS.GT.3.8 AND.BITS.LT.4.2) QLEVEL=5.8
            IF(BITS.GT.2.8.AND.BITS.LT.3.2) QLEVEL=2.5
            IF(BITS.GT.1.8.AND.BITS.LT.2.2) QLEVEL=1.25
            IF(BITS.GT..8 AND.BITS.LT.1.2) QLEVEL=0.625
            IF(BITS.GT.5.8) QLEVEL=10.8
            IF(BITS.LT.1.8) QLEVEL=0.625
      C
      C          CALL DECODER
      C
            CALL UNCODE(QLEVEL, BITS, KODE, N, SIGNAL, IKODE)
            WRITE(2) SIGNAL
            WRITE(5, 100) (SIGNAL(K), K=1, N)
      100   FORMAT(1H ,'DECODED SIGNAL ', 8F7.3)
            IF(II.EQ.5) GO TO 99
            GO TO 1
      99    STOP
            END
      C
      C
```

DECODE

MAX

```
            SUBROUTINE MAX(KODE, N, IBIG, NUM)
      C
      C          FINDS MAX IN ARRAY
      C
            IMPLICIT INTEGER*2(I-N)
            INTEGER*2 KODE(1)
            N1=N-1
            DO 20 I=1, N1
            IF(KODE(I).GT.KODE(I+1)) GO TO 10
            IBIG=KODE(I+1)
            NUM=I+1
            GO TO 20
      10    IBIG=KODE(I)
            KODE(I+1)=IBIG
      20    CONTINUE
            RETURN
            END
      C
      C
```

```
C          SUBROUTINE RESULT
C
C          ROUTINE CALCULATES SNR; WRITTEN BY L. L. BURGE
C
      IMPLICIT INTEGER*2(I-N)
      INTEGER*2 NFU(4),IBW(5)
      DIMENSION SIGNAL(256),SREAL(256),SIMAG(256),REM(256)
      DIMENSION EXCIT(256),YEXCIT(256),SNR1(258),RHAT(256)
      DIMENSION PHA2(256),SMAG(256),SPECTM(256),REIM(256)
      DIMENSION SMAGSQ(256),X(256),Y(256),TITLE2(2)
      DIMENSION BUFF(56),DATA(258),TIME(258),TITLE(2)
      EQUIVALENCE (PHA2(1),SNR1(1),DATA(1))
      EQUIVALENCE (SMAG(1),X(1),REM(1))
      EQUIVALENCE (SIGNAL(1),SREAL(1))
      EQUIVALENCE (SMAGSQ(1),SPECTM(1))
      EQUIVALENCE (EXCIT(1),RHAT(1))
      EQUIVALENCE (SIMAG(1),REIM(1),TIME(1))
      EQUIVALENCE (YEXCIT(1),Y(1))
      DATA NFU/224,214,212,192/,IBW/6,26,52,56,106/
      DATA TITLE/'HVB ','SNR='/,TITLE2/'CUM ','SNR='/
      DATA YES/'YES '/
      TWOPI=6.2832
      NB=8
      N=256
      I=8
      RSNR=0.0
      FREQ=31.25
      IBAND=1
      ANG=TWOPI/12.0
      DO 5 I=1,N
      EXCIT(I)=0.0
5     YEXCIT(I)=0.0
C
C          READ DATA
C
      REWIND 1
      REWIND 2
      REWIND 3
      REWIND 4
      REWIND 7
1     READ(1,END=99) SIGNAL
      WRITE(6,998)
998   FORMAT(1H1,' ***** ROUTINE RESULT PROCESSING ******
     *******  OUTPUT SIGNAL-TO-NOISE RATIO***** ')
      WRITE(6,938) (SIGNAL(I),I=1,N)
938   FORMAT(1H ,' INPUT SIGNAL ',8F7.3)
C
C          INTERPOLATING SIGNAL TO UP SAMPLE
C
      WRITE(6,908)
908   FORMAT(1H1,'PROCESSING SIGNAL ')
      NN=TWOPI*FREQ*IBW(IBAND)
      CALL FFFT(SREAL,SIMAG,NB,L,1)
      DO 10 I=1,N
      SMAGSQ(I)=SREAL(I)*SREAL(I)+SIMAG(I)*SIMAG(I)
10    SMAG(I)=SQRT(SMAGSQ(I))
      WRITE(6,948) (SREAL(I),SIMAG(I),SMAGSQ(I),SMAG(I),I=1,128)
948   FORMAT(1H ,'REAL  IMAG  MAGSQ  MAG  ',4F11.3)
      MOVE=IFU(IBAND)
      ISTOP=IBAND+1
      IZ=IBW(ISTOP)
      IZ2=N-IZ
      DO 38 J=1,128
      J1=MOVE+J-1
      SREAL(J)=SREAL(J1)
      SIMAG(J)=SIMAG(J1)
      J2=N-J+1
      SREAL(J2)=SREAL(J)
      SIMAG(J2)=SIMAG(J)
38    CONTINUE
      DO 40 I=IZ,IZ2
      SREAL(I)=0.0
      SIMAG(I)=0.0
40    CONTINUE
      DO 45 I=1,N
45    SMAG(I)=SREAL(I)*SREAL(I)+SIMAG(I)*SIMAG(I)
      WRITE(6,915) MOVE,IZ,ISTOP,IZ2
      WRITE(6,925) (SMAG(K),K=1,N)
915   FORMAT(1H ,4I6)
925   FORMAT(1H ,'FREQ INT SIGNAL',2F12.4)
C
C          MODULATING SIGNAL BY COSINE WAVE
C
      WRITE(6,918)
918   FORMAT(1H1,'MODULATING SIGNAL ')
      DO 50 K=1,N
50    REM(K)=COS(NN*K*ANG)
      CALL FFFT(REM,REIM,NB,L,1)
      DO 55 K=1,N
      SREAL(K)=REM(K)
      SIMAG(K)=REIM(K)
      REIM(K)=REM(K)*REM(K)+REIM(K)*REIM(K)
```

RESULT

```
 93  55     REM(K)=REIM(K)*SMAG(K)
 94         WRITE(6,950) (REM(K),K=1,N)
 95  950    FORMAT(1H ,'  MODULATED SIGNAL.   ',8F7.3)
 96         IB=IBAND+1
 97         IBA=IPHC(IPHD)
 98         IWN1=IEH(11)
 99         DO 60 K=IBA,IWN1
100         EXCIT(K)=SPEAK(K)
101         YEXCIT(K)=SIMAG(K)
102  60     CONTINUE
103         WRITE(6,960) (EXCIT(K),YEXCIT(K),K=1,IWN1)
104  960    FORMAT(1H ,'EX SIG',8F9.3)
105         IBAND=IBAND+1
106         IF(IBAND.NE.5) GO TO 1
107  99     CONTINUE
108         CALL FFFT(EXCIT,YEXCIT,NB,L,-1)
109         DO 150 I=1,N
110         EXCIT(I)=EXCIT(I)/256.0
111  150    YEXCIT(I)=YEXCIT(I)/256.0
112         WRITE(6,980) (EXCIT(K),YEXCIT(K),K=1,N)
113  980    FORMAT(1H ,'  TIME   SIGNAL ',4E15.3)
114  C
115  C         READ IN ACTUAL SIGNAL.
116  C
117         READ(3) SPECTM
118         READ(4) PHAZ
119         DO 80 K=1,N
120         SMAG(K)=SQRT(SPECTM(K))
121         X(K)=SMAG(K)*COS(PHAZ(K))
122         Y(K)=SMAG(K)*SIN(PHAZ(K))
123  80     CONTINUE
124         WRITE(6,970) (X(K),Y(K),K=1,N)
125  970    FORMAT(1H ,'  REAL   IMAG  ',4F11.3)
126         DO 90 K=129,N
127         X(K)=0.0
128         Y(K)=0.0
129  90     CONTINUE
130         WRITE(6,920)
131  920    FORMAT(1H1,'CALCULATING SNR ')
132         CALL FFFT(X,Y,NB,L,-1)
133         DO 140 K=1,N
134         X(K)=X(K)/256.0
135  140    Y(K)=Y(K)/256.0
136         WRITE(7) X
137         WRITE(6,980) (X(K),Y(K),K=1,N)
138  C
```

RESULT

```
139  C         CALL SNR ROUTINE,CALCULATE AVR SNR
140  C
141         CALL SNR(XHAT,X,N,SNR1,SGNR)
142         DO 130 K=1,N
143         TIME(K)=FLOAT(K)
144         ASNR=SNR1(K)+ASNR
145  130    CONTINUE
146         ASNR=ASNR/FLOAT(N)
147         WRITE(2) SNR1
148  C
149  C         PLOT ROUTINE
150  C
151         WRITE(5,160)
152  100    FORMAT('DO YOU WANT A SGNR PLOT')
153         READ(5,110) ANS
154  110    FORMAT(1A4)
155         IF(ANS.NE.YES) GO TO 999
156         WRITE(5,120)
157  120    FORMAT('INPUT SOUND NAME')
158         READ(5,110) CHAR
159         CALL PLOTS(BUFF,200,0)
160         CALL PLOT(0.0,-11.0,-3)
161         CALL PLOT(2.0,7.0,3)
162         CALL SYMBOL(2.0,0.0,.14,'PHONEME ',0.0,8)
163         CALL SYMBOL(999.0,999.0,.14,CHAR,0.0,4)
164         CALL SYMBOL(999.0,999.0,.14,'SGNR ',0.0,5)
165         CALL SYMBOL(3.0,7.5,.14,TITLE,0.0,8)
166         CALL NUMBER(999.0,999.0,.14,ASNR,0.0,2)
167         CALL SYMBOL(3.0,6.5,.14,TITLE2,0.0,8)
168         CALL NUMBER(999.0,999.0,.14,SGNR,0.0,2)
169         CALL PLOT(0.0,1.0,-3)
170         CALL SCALE(DATA,6.0,256,1)
171         CALL SCALE(TIME,7.0,256,1)
172         CALL AXIS(0.0,0.0,13HTIME SAMPLES ,-13,7.0,0.0,
173        +TIME(257),TIME(258))
174         CALL AXIS(0.0,0.0,21HSIGNAL TO NOISE RATIO,21,6.0,90.0,
175        +DATA(257),DATA(258))
176         CALL LINE(TIME,DATA,256,1,0,0)
177         CALL PLOT(9.0,0.0,-999)
178  999    STOP
179         END
180  C
181  C
```

SAMPLE

```
      SUBROUTINE SAMPLE(DEMOD,FASE,IBAND,I,NB,L,DEC)
C
C        COMPUTES IDEALIZED DECIMATION FUNCTION
C
      IMPLICIT INTEGER*2(I-N)
      INTEGER*2 IBAND(1),NFN(4)
      DIMENSION DEMOD(1),FASE(1),DEC(1),PHAZE(256),X(256),Y(256)
      DATA NFN/27,53,57,105/
      J=0
      IFN=NFN(I)
      WRITE(5,110)
110   FORMAT(1H1,'ROUTINE SAMPLE PROCESSING ')
      DO 10 K=1,256
      PHAZE(K)=FASE(K)
10    DEC(K)=DEMOD(K)
      DO 20 K=IFN,256
      PHAZE(K)=0.0
20    DEC(K)=0.0
      NFNA=IFN-1
      NST=NFNA/2
      DO 30 K=NST,NFNA
      J1=256-NST+1
      PHAZE(K)=FASE(J1)
      DEC(K)=DEMOD(J1)
      J=J+1
30    CONTINUE
      WRITE(5,120)
      WRITE(5,130) (DEC(K),PHAZE(K),K,K=1,256,32)
120   FORMAT(1H ,'SHIFTED SIGNAL BEFORE DECIMATION')
130   FORMAT(1H ,2F10.4,I6)
      DO 40 K=1,256
      X(K)=DEC(K)*COS(PHAZE(K))
40    Y(K)=DEC(K)*SIN(PHAZE(K))
      L=7
      CALL FFFT(X,Y,NB,L,-1)
      L=8
      DO 50 K=1,256
50    DEC(K)=X(K)
      WRITE(5,100) (DEC(K),K=1,256)
100   FORMAT(1H ,'DECIMATED SIGNAL',8F6.3)
      RETURN
      END
```

SPLIT

```
      SUBROUTINE SPLIT(AVES,PHASE,IBW,I,NB,L,DEMOD)
C
C        COMPUTES IDEALIZED BANDPASS FILTER
C
      IMPLICIT INTEGER*2(I-N)
      INTEGER*2 IBW(1)
      DIMENSION AVES(1),PHASE(1),DEMOD(1)
      DIMENSION BP(256),X(256),Y(256)
      DIMENSION PH(256)
      TWOPI=6.283185
      HN=TWOPI*IBW(I)*31.25
      WRITE(5,100)
100   FORMAT(1H0,'ROUTINE SPLIT PROCESSING')
      J=1
      I1=I+1
      DO 10 K=1,256
      PH(K)=0.0
      BP(K)=0.0
10    CONTINUE
      IPASS=IBW(I)
      IZERO=IBW(I1)
      DO 20 K=1,PASS,IZERO
      PH(J)=PHASE(K)
      BP(J)=AVES(K)
      J=J+1
20    CONTINUE
      J1=256-J
      IPASS1=256-IPASS
      IZERO1=256-IZERO
      DO 30 K=IZERO1,IPASS1
      BP(J1)=AVES(K)
      PH(J1)=PHASE(K)
      J1=J1+1
30    CONTINUE
      DO 40 K=1,256
      X(K)=BP(K)*COS(PH(K))
40    Y(K)=BP(K)*SIN(PH(K))
      CALL FFFT(X,Y,NB,L,-1)
      DO 50 K=1,256
50    DEMOD(K)=X(K)*COS(HN*K)/256.0
      WRITE(5,110) (DEMOD(K),K=1,8)
110   FORMAT(1H ,'MODULATED SIGNAL  ',8F7.2)
      RETURN
      END
C
C
```

```
C            CALCULATION OF PITCH PERIOD ESTIMATION
C          SUBROUTINE PITCH
C          CALCULATION OF PITCH PERIOD ESTIMATION BY L. L. BURGE
C          NT - VECTOR MARKING PITCH PULSE OCCURRENCES
C          IDIFF - DISTANCES BETWEEN PULSES IN SAMPLE POINTS
C          RO - PREDICTION RESIDUAL SAMPLE POINTS
C          SQR - VECTOR CONTAINING SQUARE OF AMPLITUDE FOR POINTS
C          T - TEMPORARY VECTOR FOR STORAGE
C          PT - VECTOR OF PITCH PERIOD VALUES
C          SPECTM - VECTOR CONTAINING MAGNITUDE SQUARE SPECTRUM PTS
         IMPLICIT INTEGER*2(I-N)
         INTEGER*2 NT(256),IDIFF(256)
         DIMENSION RO(256),SQR(256),T(256),PT(256)
         DIMENSION SPECTM(256)
         AVP=0.0
         N=256
         IBLK=0
         CON=31.25
         DO 1 I=1,N
         IDIFF(I)=0.0
         PT(I)=0.0
1        NT(I)=0
         TIME=0.000125
         REWIND 1
         REWIND 2
         REWIND 3
C                READ SPECTRUM DATA
         READ(1) SPECTM
         DO 20 I=1,11
         IF(SPECTM(I).GT.SPECTM(I+1)) GO TO 30
         MAXSP=SPECTM(I+1)
         FREQ=CON*FLOAT(I+1)
         GO TO 20
30       MAXSP=SPECTM(I)
         SPECTM(I+1)=MAXSP
20       CONTINUE
         PITCH=1./FREQ
         WRITE(6,900) PITCH,FREQ
900      FORMAT(1HA,'ESTIMATE OF PITCH PERIOD=',F10.4,
     * AND FUNDAMENTAL=',F10.4,' AS COMPUTED FROM MAG SPEC A DATA')
C                READ RESIDUAL DATA
100      READ(2,END=99) RO
         IBLK=IBLK+1
C              SQUARE DATA
         DO 40 M=1,N
         SQR(M)=RO(M)*RO(M)
40       SQR(M)=SQR(M)+SQR(M)
         WRITE(3) SQR
         WRITE(6,940) IBLK
         DO 50 I=1,N
50       T(I)=SQR(I)
         NP=N/4
         AVG=PITCH
         DO 60 L=1,N,NP
         JN=1
         J=L+NP-1
         DO 63 K=L,J
         K1=K+1
         IF(T(K1).GT.T(K1+1)) GO TO 64
         MAX=T(K1+1)
         GO TO 63
64       MAX=T(K1)
         T(K1+1)=MAX
63       CONTINUE
         WRITE(5,910) MAX
         DO 52 I=1,N
52       T(I)=SQR(I)
         THR=.75*MAX
         DO 69 KI=1,N
69       NT(KI)=0
         DO 65 K=L,J
         IF(T(K).GE.THR) GO TO 66
         GO TO 65
66       NT(JN)=K
         JN=JN+1
65       CONTINUE
         DO 67 I=1,JN
         IDIFF(I)=IABS(NT(I)-NT(I+1))
67       PT(I)=FLOAT(IDIFF(I))*TIME
         PTC=0.0
         DO 70 KL=1,JN
70       PTC=PT(KL)+PTC
         PTC=PTC/FLOAT(JN)
         IF(PTC.LT..001) GO TO 68
         IF(PTC.GT..02) GO TO 68
         DO 68 I=1,JN
68       AVP=PT(I)+AVP
         AVP=AVP/FLOAT(JN)
         F0=1./AVP
         IF(F0.GT.1000.) GO TO 68
         IF(F0.LT.100.) AVP=.008
         F0=1/AVP
```

PITCH

PITCH

```
      JNS=JN-1
      WRITE(6,920) (NT(M),M=1,JNS)
      WRITE(6,970) (PT(I),IDIFF(I),I=1,JNS)
      WRITE(6,980) AVP,F0
60    CONTINUE
910   FORMAT(1H0,'MAX USED AS THR REFERENCE=',F10.4)
920   FORMAT(1H0,'OCCURRENCES OF PITCH PULSE'/1H ,10I6)
930   FORMAT(1H0,'VALUES OF SELECTED PITCH PERIOD AND PULSE
     + DISTANCES '/6(F10.4,I6))
940   FORMAT(1H0,'DATA BLOCK NUMBER',I6)
      GO TO 100
99    STOP
      END
```

MAX

```
      SUBROUTINE MAX(X,N,BIG,NUM)
C
C        DETERMINES THE MAXIMUM OF AN ARRAY BY L. L. BUROE
C
      IMPLICIT INTEGER*2(I-N)
      DIMENSION X(1)
      N1=N-1
      DO 20 I=1,N1
      IF(ABS(X(I)).GT.ABS(X(I+1))) GO TO 10
      BIG=X(I+1)
      NUM=I+1
      GO TO 20
10    BIG=X(I)
      X(I+1)=BIG
20    CONTINUE
      BIG=ABS(BIG)
      RETURN
      END
```

SUMLPD

```
C     SUBROUTINE SUMLPD
C     PASSES PREDICTION RESIDUAL THRU DIGITAL BANDPASS FILTER
C     MODULATES, LOWPASS FILTERS AND DECIMATES WRITTEN BY L.L. BUROE
      IMPLICIT INTEGER*2(I-N)
      INTEGER*2 IBAND(5)
      DIMENSION DEMOD(256),DEC(256)
      DIMENSION AVES(256),PHASE(256),FASE(256)
      EQUIVALENCE (PHASE(1),FASE(1))
      DATA IBAND/6,23,42,64,90/
      L=8
      N=256
      NB=8
      DO 10 I=1,N
      AVES(I)=0.0
      PHASE(I)=0.0
10    CONTINUE
C
C       READ IN DATA
C
      REWIND 7
      REWIND 1
      REWIND 2
      REWIND 3
      REWIND 4
      REWIND 8
      READ(7,END=99) AVES
      READ(8) PHASE
      WRITE(5,100)
100   FORMAT(1H1,'***** PROCESSING BLOCK *****',4X,'**********')
99    DO 30 J=1,N
      AVES(J)=SQRT(AVES(J))
30    CONTINUE
C
C       PROCESSING SUBBANDS
C
      DO 40 I=1,4
      WRITE(5,110) I
110   FORMAT(1H0,'***** PROCESSING SUBBAND *****',I4,4X,'********')
      CALL SPLIT(AVES,PHASE,IBAND,I,NB,L,DEMOD)
      CALL LOWPAS(DEMOD,IBAND,1,NB,L,FASE)
      CALL SAMPLE(DEMOD,FASE,IBAND,I,NB,L,DEC)
      WRITE(1) DEC
40    CONTINUE
      STOP
      END
C
```

SNRCAL

```
      C
      C              SUBROUTINE SNRCAL
      C
      C         CALCULATE SNR FROM 256 POINT SIGNAL
      C
            IMPLICIT INTEGER*2(I-N)
            DIMENSION SIGNAL(256),SIGEST(256),QNR(256),ACTUAL(256)
            DIMENSION ERROR(256)
            N=256
            DUMSUM=0.0
            SUMSUM=0.0
            ASNR=0.0
      C
      C         READ IN DATA
      C
            REWIND 1
            REWIND 2
            READ(1,END=99) SIGNAL
            READ(2) SIGEST
      C
      C         CALCULATE CUMULATIVE SUM FOR SNR
      C
            DO 10 I=1,N
            ACTUAL(I)=SIGNAL(I)*SIGNAL(I)
            ERROR(I)=(SIGNAL(I)-SIGEST(I))*(SIGNAL(I)-SIGEST(I))
            DUMSUM=ACTUAL(I)+DUMSUM
            SUMSUM=ERROR(I)+SUMSUM
      10    CONTINUE
      C
      C         CALCULATE NORMALIZED SNR & RMS SNR
      C
            SNR=SUMSUM/DUMSUM
            RMS=SUMSUM/N
            SQNR=10*ALOG10(SNR)
            RMS=SQRT(RMS)
            RMS=20*ALOG10(RMS)
      C
      C         CALCULATE MEAN SQUARE ERROR
      C
            DO 20 I=1,N
            IF(ACTUAL(I).EQ.0.0) ACTUAL(I)=0.001
            QNR(I)=ABS(ERROR(I)/ACTUAL(I))
            IF(QNR(I).EQ.0.0) QNR(I)=0.001
            QNR(I)=10.0*ALOG10(QNR(I))
            ASNR=QNR(I)+ASNR
      20    CONTINUE
            ASNR=ASNR/FLOAT(N)
```

SNRCAL

```
      C
      C         WRITE THE CALCULATED RESULTS
      C
            WRITE(5,910) RMS,SQNR,ASNR
      910   FORMAT(1H0,' RMS SQNR ',3X,F10.3/1H ,' NORM SQNR ',3X,F10.3
     +/1H ,' MEAN SQUARE SQNR ',3X,F10.3)
      99    STOP
            END
```

UNCODE

```
            SUBROUTINE UNCODE(QLEVEL,BITS,KODE,N,SIGNAL,IKODE)
      C
      C       DECODER FOR INPUT CODE, BY L. L. BURKE
      C
            IMPLICIT INTEGER*2(I-N)
            INTEGER*2 KODE(1),IKODE(1)
            DIMENSION SIGNAL(1)
            IBITS=IFIX(BITS+.1)
            IQNL=2**IBITS
            QINC=QLEVEL/FLOAT(IQNL)
            QNEG=FLOAT(IQNL)/2.0
            IQNEG=IFIX(QNEG+.1)
            IQPOS=IQNL
            WRITE(5,100) IQNL,IBITS,QINC
      100   FORMAT(1H0,'NUMBER LEVELS',I3,2X,'BITS',I3,2X,'INC',F7.3)
            DO 10 I=1,N
            DO 20 J=1,IQNEG
            J1=IQNEG-1+J
            IF(KODE(I).EQ.IKODE(J)) SIGNAL(I)=-FLOAT(J1)*QINC
      20    CONTINUE
            K1=1
            DO 30 K=IQNEG,IQPOS
            IF(KODE(I).EQ.IKODE(K)) SIGNAL(I)=FLOAT(K1)*QINC
            K1=K1+1
      30    CONTINUE
      10    CONTINUE
            RETURN
            END
      C
      C
```

```
      SUBROUTINE FFTMGR
      IMPLICIT INTEGER*2(I-N)
C     CALCULATES THE MAGNITUDE SQUARED OF FOURIER SPECTRUM
C     WRITTEN BY L. I. BURGE
C     CALCULATES ENERGY FOR N PTS
C     S1 - VECTOR OF INPUT SAMPLES ON INPUT
C     S1 - VECTOR OF REAL SPECTRUM POINTS ON OUTPUT FROM FFT
C    S1 - VECTOR OF IMAGINARY SPECTRUM POINTS OUTPUT FROM MAGSQ
C     Y1 - VECTOR OF MAGNITUDE SPECTRUM POINTS ON OUTPUT
C     N - NUMBER OF POINTS
C     NBITS1 - ORDER OF FFT
C     AVES - SPECTRAL COEFFICIENTS
C     ESUB - SUB-BAND ENERGY
C     SIGMA - NORMALIZATION CONSTANT
      INTEGER*2 KBITS(3,4)
      DIMENSION S1(258),Y(258),Y1(258),E(256),ZBITS(4),AVEP(256)
      DIMENSION AVES(258),XP(258),DATA(258),BUFF(5A),PHAZ(256)
      DIMENSION BITS(4),ESUB(4),ENERGY(3,4),SIGMA(4),SEN(16)
      EQUIVALENCE (XP(1),Y(1),E(1)),(DATA(1),AVES(1))
      DATA YES/'YES '/
      DATA SIGMA/4*1.0/
      DATA KBITS/4,3,2,3,3,3,2,3,3,2,2,3/
      DATA ENERGY/10.,7.,4.,5.,3.,2.,
     +24.,16.,8.,16.,16.,8./
      ICOUNT = 0
      IN=0
      NBITS1=9
      DO 1 I=1,256
      AVEP(I)=0.0
1     AVES(I)=0.0
      DO 2 I = 1,4
2     ZBITS(I) = 0.0
      CON=31.25
      N=256
      NBITS = NBITS1 -1
      L = 0
C
C      READ DATA
C
      REWIND 1
      REWIND 2
      REWIND 3
      REWIND 4
5     READ(1,END=99) S1
C
C        CALCULATE SPECTM AND ENERGY
C
      CALL FFFT(S1,Y1,NBITS,L,1)
      CALL MAGSQ(Y1,S1,N,E,PHAZ)
      IN=IN+1
      BE=E(256)/2.0
      IF(IN.EQ.1) GO TO 630
      IF(E(256).LT.1.0) GO TO 610
630   WRITE(6,111) IN,BE
111   FORMAT(1H0,'SECTION =',I4,' TOTAL  ENERGY FOR BLOCK = ',F16.6)
C
C        CALCULATE ENERGY/SUB-BAND BAND
C
      CALL SUB(Y1,N,ESUB,IN)
      E11=10.0
      E22=4.0
      E33=1.0
      SEN(IN)=E(256)
C
      IF(E(256).LT.40.0) GO TO 200
C
C        CHECK IF HIGH OR LOW ENERGY BASED ON PREVIOUS ENERGY
C
      CALL CHKENG(E,SEN,IN,ESUB)
200   ETOT=E(256)
      IF(ETOT.LT.E33) GO TO 503
C
C        A PRIORI CALCULATION OF NORMALIZATION CONSTANT
C
      IF(ETOT.GT.E11) GO TO 501
      IF(ETOT.GT.E22) GO TO 502
      IF(ETOT.GT.E33) GO TO 503
501   DO 511 J=1,4
511   SIGMA(J)=ENERGY(1,J)/(10.**(KBITS(1,J)/3.32193))
      GO TO 540
502   DO 522 J=1,4
522   SIGMA(J)=ENERGY(2,J)/(10.**(KBITS(2,J)/3.32193))
      GO TO 540
503   DO 533 J=1,4
      SIGMA(J)=ENERGY(3,J)/(10.**(KBITS(3,J)/3.32193))
      IF(SIGMA(J).EQ.0.0) SIGMA(J)=1.0
533   CONTINUE
540   CONTINUE
      WRITE(6,550) (SIGMA(J),J=1,4)
550   FORMAT(1H0,'SIGMA VALUES =',4F16.6)
555   FORMAT(1H0,'BITS/SAMPLE MATRIX'/1H0,4I8/
     +1H0,4I8/1H0,4I8)
```

FFTMGR

```
 93  560    FORMAT(1H0,'ENERGY THRESHOLD MATRIX'/
 94     *  +1H0,4F16.6/1H0,4F16.6/1H0,4F16.6)
 95  C
 96  C         CALCULATE BITS FOR SPECIFIED SIGMA
 97  C
 98         DO 606 I=1,4
 99  606    BITS(I)=3.32193*ALOG10(ESUB(I)/SIGMA(I)+1.0)
100         WRITE(6,605) (BITS(I),I=1,4)
101         TBITS=0.0
102         DO 601 I=1,4
103  601    TBITS=BITS(I)+TBITS
104         IF(TBITS.LT.2.0) GO TO 603
105         ICOUNT=ICOUNT+1
106         DO 602 I=1,4
107  602    ZBITS(I)=ZBITS(I)+BITS(I)
108  603    CONTINUE
109  605    FORMAT(1H0,'BITS ALLOCATED FOR SUBBANDS',2X,4F12.3)
110         GO TO 651
111  610    WRITE(6,620) IN
112  620   FORMAT(1H0,'SECTION',I4,'     IS PREDOMINATELY NOISE',
113        *'******** NO ENERGY CALCULATED')
114  C
115  C         ESTIMATION OF AVERGE SPECTRUM AND AVERAGE BITS
116  C
117  651    DO 10 I=1,256
118         AVES(I)=AVES(I)+Y1(I)
119         AVEP(I)=AVEP(I)+PHAZ(I)
120  10     CONTINUE
121         GO TO 5
122  99     DO 20 I=1,256
123         AVES(I)=AVES(I)/16.0
124         AVEP(I)=AVEP(I)/16.0
125  20     CONTINUE
126         DO 664 I=1,4
127  664    ZBITS(I)=ZBITS(I)/16.0
128         DO 30 I=1,256
129         XP(I)=(I-1)*CON
130  30     CONTINUE
131         WRITE(6,555) ((KBITS(I,J),J=1,4),I=1,3)
132         WRITE(6,560) ((ENERGY(I,J),J=1,4),I=1,3)
133         WRITE(6,565) (ZBITS(I),I=1,4)
134  565    FORMAT(1H0,'AVERAGED BITS ALLOCATED TO SUBBAND',2X,4F6.3)
135  C
136  C         WRITE OUT DATA ON DISK AND PRINTER
137  C
138         WRITE(2) AVES
```



```
139         WRITE(3) ZBITS
140         WRITE(4) AVEP
141  C
142  C         PLOTTING ROUNTINE
143  C
144         WRITE(5,112)
145  112    FORMAT('DO YOU WANT A PLOT?')
146         READ(5,113) ANS
147  113    FORMAT(1A4)
148         IF(ANS.NE.YES ) GO TO 999
149         WRITE(5,130)
150  130    FORMAT('INPUT SOUND NAME')
151         READ(5,113) CHAR
152         CALL PLOTS(BUFF,206,0)
153         CALL PLOT(0.0,-11.0,-3)
154         CALL PLOT(3.0,7.0,3)
155         CALL SYMBOL(2.0,7.0,.14,'PHONEME ',0.0,8)
156         CALL SYMBOL(999.0,999.0,.14,CHAR,0.0,4)
157         CALL PLOT(0.0,1.0,-3)
158         CALL SCALE(DATA,6.0,256,1)
159         CALL SCALE(XP,6.0,256,1)
160         CALL AXIS(0.0,0.0,9HFREQUENCY,-9,6.0,0.0,XP(257),XP(258))
161         CALL AXIS(0.0,0.0,11HMAG SQUARED,11,6.0,90.0,DATA(257),DATA(258)),0.
162         CALL LINE(XP,AVES,256,1,0,0)
163         CALL PLOT(10.0,0.0,-999)
164  999    STOP
165         END
166  C
167  C
```

MAGSQ

```
      SUBROUTINE MAGSQ(X,Y1,NT,E,PHAZ)
      IMPLICIT INTEGER*2(I-N)
C      PROGRAM COMPUTES MAGNITUDE SQUARED OF TWO ARRAYS
C      WRITTEN BY L. L. BURGE
C      X - FIRST ARRAY FOR INPUT AND MAGNITUDE ARRAY FOR OUTPUT
C      Y1 - SECOND ARRAY FOR INPUT
C      NT - NUMBER OF POINTS
C        E - ENERGY
      DIMENSION X(1),Y1(1),E(1),PHAZ(1)
      DO 10 J = 1,NT
      PHAZ(J)=ATAN2(X(J),Y1(J))
10    X(J)=X(J)*X(J)+Y1(J)*Y1(J)
      CALL ENERGY(X,NT,E)
      RETURN
      END
C
C
```

ENERGY

```
      SUBROUTINE ENERGY(X,NT,E)
      IMPLICIT INTEGER*2(I-N)
C      THIS PROGRAM CALCULATES THE ENERGY PER N POINTS
C      WRITTEN BY L. L. BURGE
C      X - MAGNITUDE SQUARED INPUT ARRAY
C      NT - NUMBER OF POINTS
C        E - ENERGY
      DIMENSION X(1),E(1)
      L = NT - 1
      E(1) = 0.0
      DO 10 J = 1,L
      E(J+1) = (0.0002444)*X(J)+E(J)
10    CONTINUE
      RETURN
      END
```

LATTIC

```
      IMPLICIT INTEGER*2(I-N)
C      IMPLEMENTATION OF LATTICE FILTER BY L. L. BURGE
C      AJ, S1 - INPUT OF SPEECH SAMPLES
C      N - NO. OF POINTS
C      M - ORDER OF FILTER
C      RC - REFLECTION COEFFICIENTS
C      RO - PREDICTION RESIDUAL OF SPEECH PTS OUTPUT
C      B - VECTOR OF BACKWARD PREDICTED SAMPLES
C      C - TEMPORARY VECTOR FOR SAMPLES
      DIMENSION R(11),A1(256),A(11),ALPHA(11),RC(11),ARRAY(44)
      DIMENSION C(11,256),B(11),SI(11)
      DIMENSION RO(258)
      DIMENSION XN(256),DATA(258),DATA1(258),BUFF(50)
      EQUIVALENCE (ARRAY(1),R(1)),(ARRAY(12),A(1))
      EQUIVALENCE (ARRAY(23),ALPHA(1)),(ARRAY(34),RC(1))
      EQUIVALENCE (DATA(1),RO(1))
      EQUIVALENCE (DATA1(1),A1(1))
      DATA YES/'YES '/
      M=10
      N=256
      NS=0
      IREC=0
      CON=1./8000.
C       READ WINDOWED DATA
      REWIND 1
      REWIND 2
      REWIND 3
5     READ(1,END=99) A1
      REWIND 5
      IREC=IREC+1
      DO 10 I=1,N
10    XN(I)=(I-1)*CON
      READ(2) ARRAY
      A(1) = 0.0
      DO 40 K = 1,N
      C(1,K)=A1(K)
      IF(K.GT.1) B(1)=A1(K-1)
      DO 30 J=1,M
      SI(1)=A1(K)
      SI(J+1)=SI(J)+RC(J)*B(J)
      C(J+1,K)=B(J)+RC(J)*SI(J)
      IF (K.EQ.1) GO TO 20
      B(J+1)=C(J+1,K-1)
      GO TO 30
20    B(J+1)=0.0
30    RO(K)=SI(J+1)
```

```
 30     CONTINUE
 40     CONTINUE
        WRITE(3) M0
        M1=1+50
        WRITE(6,100) (RO(I),I=1,M1)
100     FORMAT(1H0,'RESIDUAL VALUES FROM LATTICE FILTER'/
     &(10E12.6))
        NS=NS+1
        WRITE(4,500) NS
C           PLOTTING ROUTINE
        WRITE(5,110)
110     FORMAT('IS A PLOT DESIRED?')
        READ(5,120) ANS
120     FORMAT(1A4)
        IF(ANS.NE.YES) GO TO 5
        WRITE(5,130)
130     FORMAT('INPUT SOUND NAME')
        READ(5,120) CHAR
        FRAME=FLOAT(IFEC)
        CALL PLOTS(BUFF,200,0)
        CALL PLOT(0.0,-11.0,-3)
        CALL PLOT(-1.0,3.0,3)
        CALL SYMBOL(-1.0,3.0,.14,'PHONEME ',90.0,8)
        CALL SYMBOL(999.0,999.0,.14,CHAR,90.0,4)
        CALL SYMBOL(999.0,999.0,.14,'FRAME ',90.0,6)
        CALL NUMBER(999.0,999.0,.14,FRAME,90.0,-1)
        CALL PLOT(6.0,1.0,-3)
        CALL SCALE(DATA,4.0,256,1)
        CALL SCALE(XN,4.0,256,1)
        CALL AXIS(0.0,0.0,13HTIME INTERVAL,-13,4.0,0.0,
     +XN(257),XN(258))
        CALL AXIS(0.0,0.0,8HRESIDUAL,8,4.0,90.0,
     +DATA(257),DATA(258))
        CALL LINE(XN,RO,256,1,0,0)
500     FORMAT(1H0,'PRESENTLY PROCESSING DATA THRU LATTICE FILTER
     +  '/1H ,'  FOR SPEECH  SECTION ',I5)
        CALL SCALE(DATA1,4.0,256,1)
        CALL AXIS(0.0,5.0,13HTIME INTERVAL,-13,4.0,0.0,
     +XN(257),XN(258))
        CALL AXIS(0.0,5.0,6HSPEECH,6,4.0,90.0,
     +DATA1(257),DATA1(258))
        CALL PLOT(0.0,5.0,-3)
        CALL LINE(XN,A1,256,1,0,0)
        CALL PLOT(6.0,0.0,-3)
        GO TO 5
99      CALL PLOT(0.0,0.0,-999)
```



```
        STOP
        END
```

SNR

```
      SUBROUTINE SNR(REST,RACT,N,QNR,SQNR)
C
C        CALCULATES SNR WRITTEN BY L. L. BURGE
C
      IMPLICIT INTEGER*2(I-N)
      DIMENSION REST(1),RACT(1),QNR(1),ACTUAL(256),ERROR(256)
      DUMSUM=0.0
      SUMDUM=0.0
      DO 20 I=1,N
      ACTUAL(I)=RACT(I)*RACT(I)
      ERROR(I)=(RACT(I)-REST(I))*(RACT(I)-REST(I))
      DUMSUM=ACTUAL(I)+DUMSUM
      SUMDUM=ERROR(I)+SUMDUM
20    CONTINUE
      SQNR=SUMDUM/DUMSUM
      SQNR=10*ALOG10(SQNR)
      DO 10 I=1,N
      IF(RACT(I).EQ.0.0) RACT(I)=0.032
      ACTUAL(I)=RACT(I)*RACT(I)
      ERROR(I)=(RACT(I)-REST(I))*(RACT(I)-REST(I))
      QNR(I)=ABS(ERROR(I)/ACTUAL(I))
      IF(QNR(I).EQ.0.0) QNR(I)=0.001
      QNR(I)=10.0*ALOG10(QNR(I))
10    CONTINUE
      RETURN
      END
C
C
```

SUB

```
      SUBROUTINE SUB(V1,N,ESUB,IN)
C       V1 - VECTOR OF MAGNITUDE SPECTRUM POINTS
C       N - NUMBER OF POINTS
C       ESUB - VECTOR OF SUB-BAND ENERGY OF OUTPUT
C       IN - FRAME NUMBER
C       THIS ROUTINE ESTIMATES ENERGY/SUB-BAND BY L.L.MURDIE
      IMPLICIT INTEGER*2(I-N)
      DIMENSION V1(1),ESUB(4)
      ESUB1=0.0
      ESUB2=0.0
      ESUB3=0.0
      ESUB4=0.0
      PTS=4096
      IFIL1=23
      IFIL2=42
      IFIL3=64
      IFIL4=90
C
C         ESTIMATION OF NORMLIZED SUB-BAND DISTRIBUTIONS
C
      DO 10 I=1,IFIL1
10    ESUB1=(1.0/PTS)*V1(I)+ESUB1
      ESUB1=2.*ESUB1
      ESUB(1)=ESUB1
      DO 20 I=IFIL1,IFIL2
20    ESUB2=(1.0/PTS)*V1(I)+ESUB2
      ESUB2=2.*ESUB2
      ESUB(2)=ESUB2
      DO 30 I=IFIL2,IFIL3
30    ESUB3=(1.0/PTS)*V1(I)+ESUB3
      ESUB3=2.*ESUB3
      ESUB(3)=ESUB3
      DO 40 I=IFIL3,IFIL4
40    ESUB4=(1.0/PTS)*V1(I)+ESUB4
      ESUB4=2.*ESUB4
      ESUB(4)=ESUB4
      WRITE(6,100) IN,(ESUB(I),I=1,4)
100    FORMAT(1H0,'SUB-BAND ENERGY DISTRIBUTED FOR SECTIONS',
     +I4/1H0,4F16.6)
      RETURN
      END
C
C
C
```

FFT

```
      SUBROUTINE FFFT(X,V1,M,L,ITRAN)
C
      IMPLICIT INTEGER*2(I-N)
C     THIS ROUTINE IS A REVISION OF
C      FAST FOURIER TRANS WITH PRUNING BY J.D. MARKEL
C      X - VECTOR OF REAL DATA POINTS
C      V1 - VECTOR OF IMAGINARY DATA POINTS
C      M - ORDER OF FFT.........2**M=NO. OF PTS.
C      L - ORDER OF PRUNED DATA (NC2**L)  LCM
      DIMENSION X(1),V1(1)
      N = 2**M
      L2 = 2**L
C
C         ZERO THE VECTOR FOR IMAGINARY POINTS
C
      DO 10 I = 1,256
10    V1(I) = 0.0
      TWOPI = 6.283185
      DO 40 LO = 1,M
      LMX = 2**(M-LO)
      LMM = LMX
      LIX = 2*LMX
      SCL1 = TWOPI/LIX
      IF(LO-M+L) 20,30,30
20    LMM = L2
C
C         ITRAN IS THE DIRECTION OF THE TRANSFORM
C
30    DO 40 LM = 1,LMM
      ARG = (LM-1)*SCL1*ITRAN
      C = COS(ARG)
      S = SIN(ARG)
      DO 40 LI = LIX,N,LIX
      J1 = LI-LIX+LM
      J2 = J1+LMX
      T1 = X(J1)-X(J2)
      T2 = V1(J1)-V1(J2)
      X(J1) = X(J1)+X(J2)
      V1(J1) = V1(J1)+V1(J2)
      X(J2) = C*T1+S*T2
40    V1(J2) = C*T2-S*T1
      CALL RBT(X,V1,M)
      RETURN
      END
C
C
```

RBT

```
      SUBROUTINE RBT(X,Y1,M)
      IMPLICIT INTEGER*2(I-N)
C      ROUTINE REVERSES BITS,0 VIS 1,VICE VERSA,DUE TO FFT
C      A REVISION OF J. D  MARKEL ROUTINE
C      X - REAL FFT POINTS
C      Y1 - IMAGINARY FFT POINTS
C      M - ORDER OF FFT
      INTEGER*2 L(9)
      DIMENSION X(1),Y1(1)
      DO 20 J=1,8
      L(J) = 1
      IF(J-M) 10,10,20
   10 L(J) = 2**(M+1-J)
   20 CONTINUE
      JN = 1
      L8 = L(8)
      DO 50 J8 = 1,L8
      L7 = L(7)
      DO 50 J7 = J8,L7,L8
      L6 = L(6)
      DO 50 J6 = J7,L6,L7
      L5 = L(5)
      DO 50 J5 = J6,L5,L6
      L4 = L(4)
      DO 50 J4 = J5,L4,L5
      L3 = L(3)
      DO 50 J3 = J4,L3,L4
      L2 = L(2)
      DO 50 J2 = J3,L2,L3
      L1 = L(1)
      DO 50 J1 = J2,L1,L2
      IF(JN-J1) 30 ,30,40
   30 R = X(JN)
      X(JN) = X(J1)
      X(J1) = R
      FI = Y1(JN)
      Y1(JN) = Y1(J1)
      Y1(J1) = FI
   40 JN = JN + 1
   50 CONTINUE
      CALL USCRM(X,Y1,M)
      RETURN
      END
C
C
C
```

USCRM

```
      SUBROUTINE USCRM(X,Y1,M)
      IMPLICIT INTEGER*2(I-N)
C      THIS PROGRAM UNSCRAMBLES FOURIER TRANSFORM OF REAL DATA
C      DUE TO FFT AFTER S. B. DAVIS
C      X - ARRAY OF REAL TRANSFORM POINTS
C      Y1 - ARRAY OF IMAGINARY TRANSFORM POINTS
C      M - ORDER OF TRANSFORM
      DIMENSION X(1),Y1(1)
      I = 2**M
      L02 = I/2
      PI = 3.14159265 .
      SA = PI/I
      X(1) = X(1) + Y1(1)
      Y1(1) = 0.0
      LLL = L02 + 1
      Y1(LLL) = -Y1(LLL)
      DO 10 K = 2,L02
      J = I-K+ 2
      ANG = SA*(K-1)
      C = COS(ANG)
      S = SIN(ANG)
      AA = X(K) + X(J)
      BB = Y1(K) - Y1(J)
      CC = X(K) - X(J)
      DD = Y1(K) + Y1(J)
      XR = C*DD-S*CC
      XI = S*DD+C*CC
      X(K) = (AA+XR)*.5
      X(J) = (AA - XR)*.5
   10 Y1(J) = (-BB-XI)*.5
      RETURN
      END
C
C
C
C
```

APPENDIX C

ARTICULATION INDEX

ARTICULATION INDEX

The concept of the Articulation Index (AI) was advanced by French and Steinberg [86]. It is defined as a number obtained from articulation tests using nonsense syllables under the assumption that any narrow band of speech frequencies of a given intensity in the absence of noise carries a contribution to the total index, which is independent of the other bands with which it is associated, and that the totals of all the bands is the sum of the contributions of the separate bands [86]. It must be proven that there is a unique function relating syllable or word articulation to AI for any given articulation crew and choice of word list. In determining AI under these conditions, there are essentially two parameters of a linear communication system that can be varied: (a) the level of the speech above the threshold of hearing, and (b) the frequency response of the system. Here a linear system that is free from noise is assumed.

A curve of AI versus frequency is included from French and Steinberg [86] in Figure 32. The curve is derived from the syllable articulation gain and frequency responses of speech waveforms [86]. The syllable articulation is expressed as the percentage of syllables with which consonant-vowel-consonant of meaningless monosyllables are perceived correctly.

Baranek [87] pointed out two important facts. First, extending the frequency range of a communication system below 200 or above 6000 Hz contributes almost nothing to the intelligibility of speech. Second,

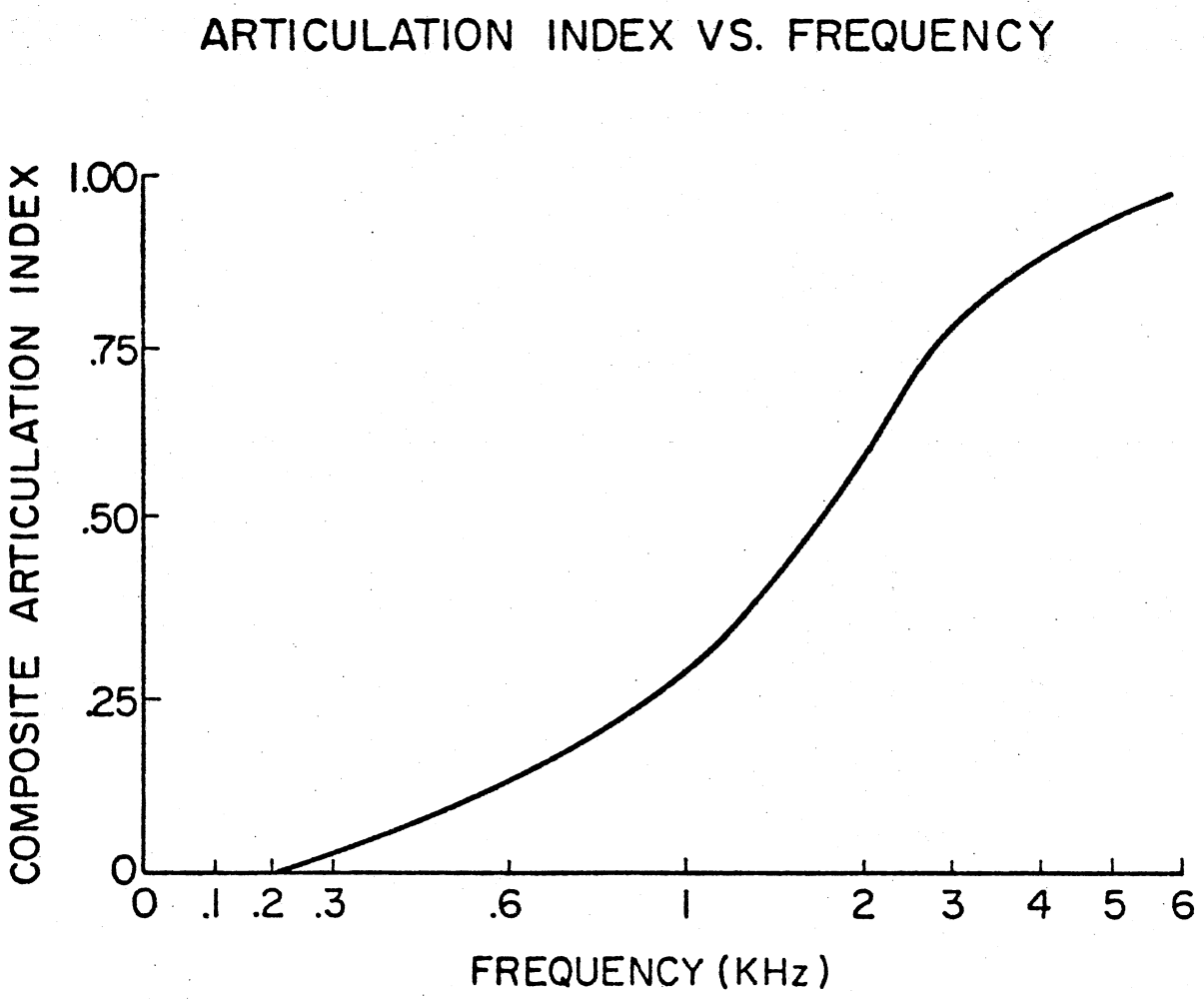


Figure 32. Composite Articulation Index vs. Cutoff Frequency of Ideal Lowpass Filters (After French and Steinberg, 1947)

each frequency band shown in Table XXV makes a 5 percent contribution to the AI, provided that the orthotelephonic gain of the system is optimal (about +10 dB) and that there is no noise present.

TABLE XXV

FREQUENCY BANDS OF EQUAL CONTRIBUTION TO ARTICULATION INDEX

No.	Edges of Band	Mid-Freq (Mean)	No.	Edges of Band	Mid-Freq (Mean)
1	200 - 300	270	11	1600 - 1830	1740
2	330 - 430	380	12	1830 - 2020	1920
3	430 - 560	490	13	2020 - 2240	2130
4	560 - 700	630	14	2240 - 2500	2370
5	700 - 840	770	15	2500 - 2820	2660
6	840 - 1000	920	16	2820 - 3200	3000
7	1000 - 1150	1070	17	3200 - 3650	3400
8	1150 - 1310	1230	18	3650 - 4250	3950
9	1310 - 1480	1400	19	4250 - 5050	4650
10	1480 - 1660	1570	20	5050 - 6100	5600

The orthotelephonic (OT) gain is defined by

$$\text{OT Gain (Subjective)} = 20 \log (e_o/p_o) + 20 \log (E_2/e_o) + 20 \log (P_1/E_2) \quad \text{(A.1)}$$

where

P_1 = free field pressure necessary to produce the same loudness in the ear as was to produce by the earphone with voltage E_2 across.

e_o = voltage produced by the microphone across the input resistor of the amplifier by a voice which produces pressure p_o at a distance of one meter in a free field.

E_2/e_o = voltage amplification of the amplifier.

$$\text{OT Gain (Objective)} = 20 \log (e_o/p_o) + 20 \log R + 20 \log (e_2/e_o) + 20 \log (p_e/e_2) \quad (A.2)$$

where

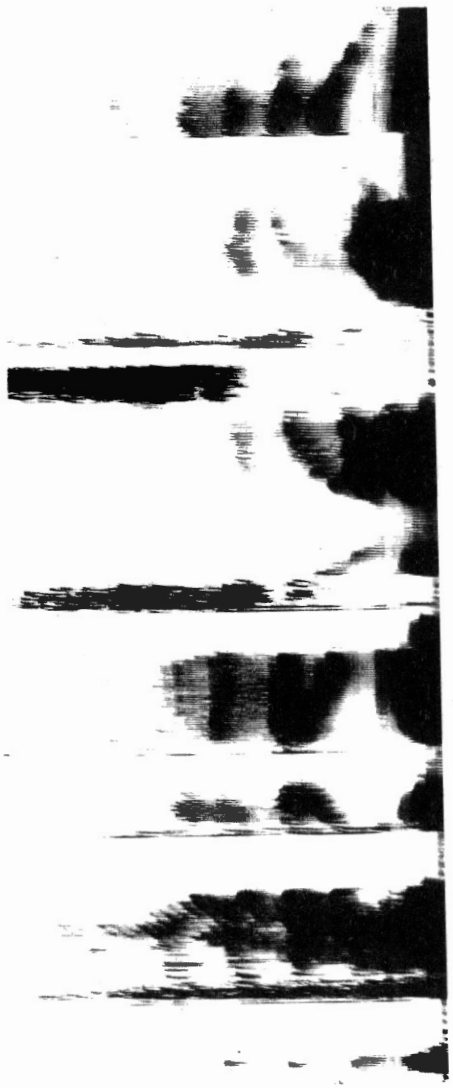
R = ratio of the pressure produced at the eardrum of a listener by a source of sound to the pressure which would be produced by the same source at the listener's head position if he were removed from the field.

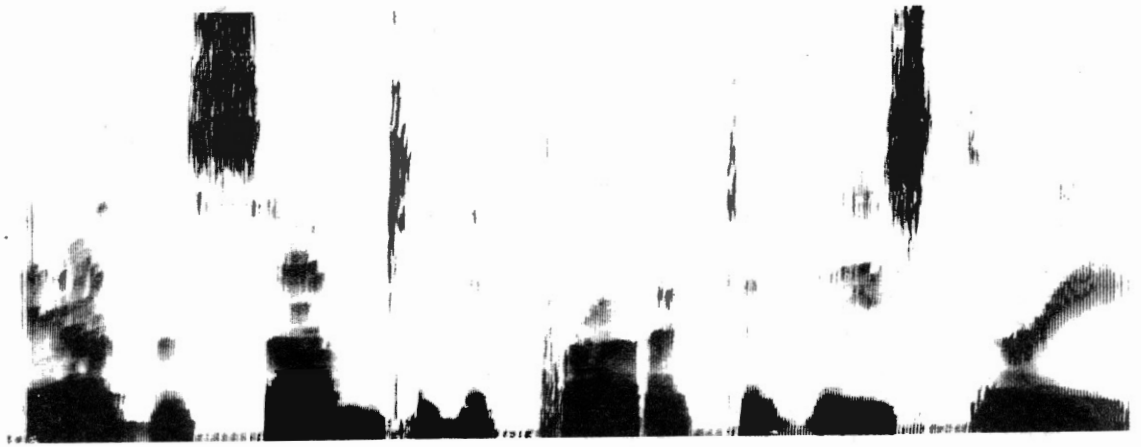
p_e = pressure produced at the eardrum of a listener by the earphone with a voltage e_2 across it; others are the same.

The AI obtained per frequency band in Table XXV is successively added to arrive at the total AI.

APPENDIX D

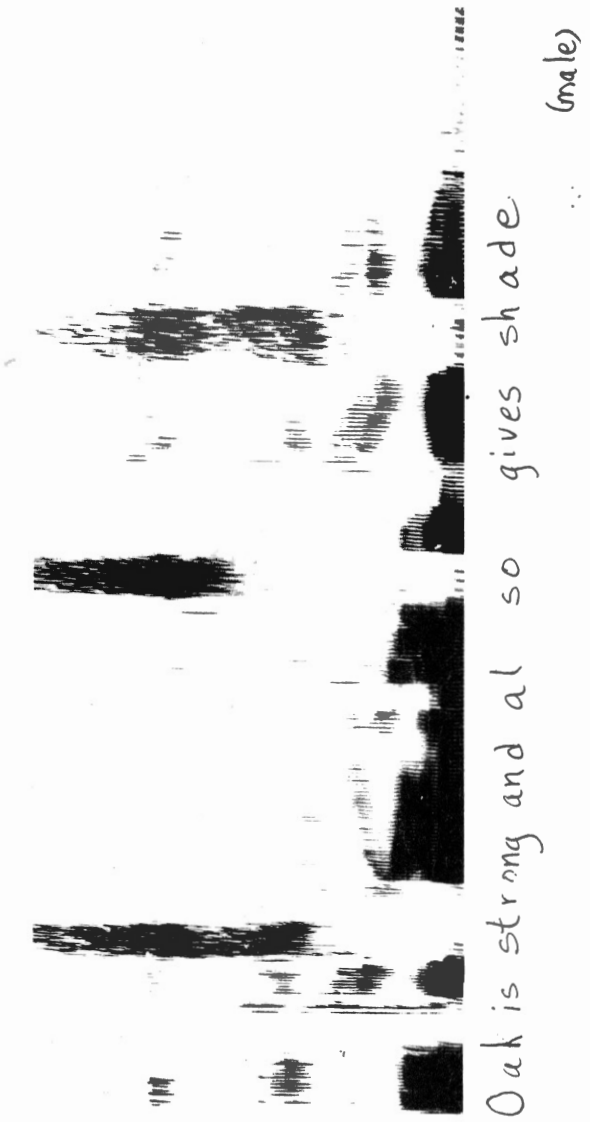
SONAGRAMS

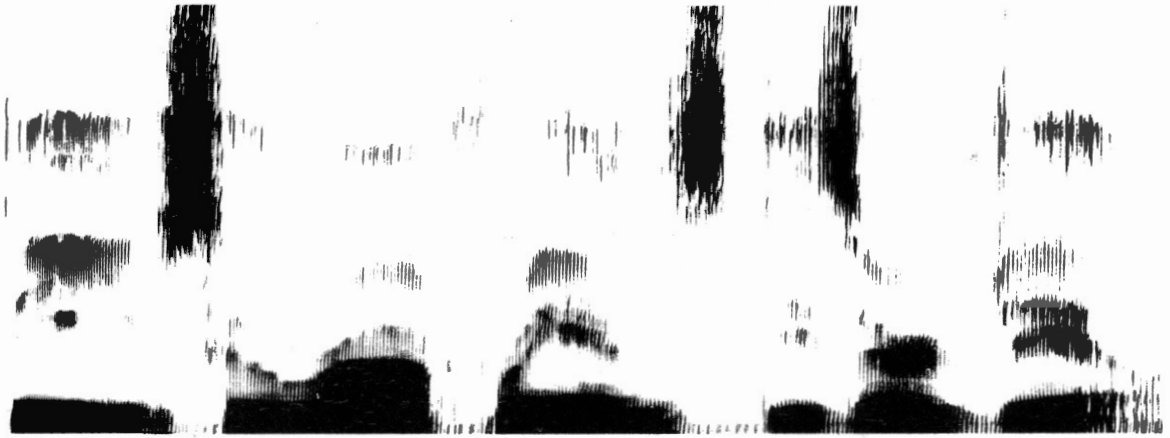

The pipe began to rust while new
(female)



Add the s u m to the pro duct of these three .
(female)


O pen the cra te but don't break the gla ss
(female)


Oak is strong and al so gives shade
(male)



Thieves who rob friends deserve jail

(male

Cat s and dog s ea ch hate the other

(male)

VITA

Legand L. Burge, Jr.

Candidate for the Degree of

Doctor of Philosophy

Thesis: EFFICIENT CODING OF THE PREDICTION RESIDUAL

Major Field: Electrical Engineering

Biographical:

Personal Data: Born in Oklahoma City, Oklahoma, August 3, 1949, the son of Mr. and Mrs. L. L. Burge.

Education: Graduated from Douglass High School, Oklahoma City, Oklahoma, in May 1967; received Bachelor of Science degree in Electrical Engineering from Oklahoma State University in 1972; received Master of Science in Electrical Engineering from Oklahoma State University in 1973; completed requirements for the Doctor of Philosophy degree at Oklahoma State University in December 1979.

Professional Experience: Engineer Trainee, Oklahoma Gas and Electric, summers 1969-71; student teaching assistant, School of Electrical Engineering, Oklahoma State University, 1970-71; graduate teaching assistant, School of Electrical Engineering, Oklahoma State University, 1971-72; United States Air Force, 1973 to present.

Professional Organizations: Member of Institute of Electrical and Electronic Engineers, Sigma Xi, Eta Kappa Nu.