Fall 1985

An Investigation to Improve Linear Predictive Vocoder Pulse/ Noise Excitation Models

Elizabeth Annella Martina Effer
Old Dominion University

AN INVESTIGATION TO IMPROVE LINEAR PREDICTIVE VOCODER PULSE/NOISE EXCITATION MODELS

by

Elizabeth Annella Martina Effer
B.S.E.E. May 1983, Old Dominion University

A Thesis Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF ENGINEERING

ELECTRICAL ENGINEERING

OLD DOMINION UNIVERSITY
December 1985

Approved by:

Stephen A. Zahorian (Director)

Jack Stoughton

Joseph Hilbey

ABSTRACT

AN INVESTIGATION TO IMPROVE LINEAR PREDICTIVE

VOCODER PULSE/NOISE EXCITATION MODELS

Elizabeth Annella Martina Effer
Old Dominion University, 1985
Director: Dr. Stephen A. Zahorian

The quality of synthetic speech from Linear Predictive (LP) vocoders is known to be degraded due to the lack of detail in the commonly used pulse/noise excitation model. In this investigation, it was hypothesized that this degradation is due to the lack of precise timing information in the pulses and to the constraint that each short-time segment of excitation be either an impulse train or white noise. Accordingly, more complex excitation models were implemented using precise timing from peaks in the residual and a mixture of pulses and noise. Since the LP residual is known to be the perfect excitation signal for LP vocoders, these models were based on the LP residual. The timing was determined by locating peaks in a lowpass-filtered LP residual energy waveform. In order to determine the approximate mixture of pulses and noise, two methods were explored to separate the periodic and non-periodic components of the residual. One method, based on the assumption

that the periodic and non-periodic components are separated in the frequency domain, employed linear filters to separate the two components. The second method, based on the assumption that the components are separated in the time domain, used time-domain techniques and the lowpass-filtered residual energy waveform to separate the two components. The time domain approach proved to be more feasible. Frequency domain models were developed for modeling the periodic pulse-like component and the non-periodic noise-like component such that the spectrum of the combined components would be flat. Listening experiments indicated that the precise timing of the periodic component resulted in improved quality synthetic speech. Improvements in speech quality related to modeling a mixture of pulses and noise in the excitation were much more difficult to obtain.

ACKNOWLEDGEMENT

I would like to thank my advisor, Dr. Stephen Zahorian, for all of the time, effort, and patience he gave me. I am truly grateful for his guidance and support.

I would like to thank the other members of my committee, Dr. Jack Stoughton, and Dr. Joseph Hibey for their help and time.

I would also like to acknowledge the support of both the Texas Instruments Fellowship and the Old Dominion University Fellowship, whose assistance made my graduate studies possible.

And to Winston and Leo for their constant love and devotion.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF SYMBOLS

SYMBOL	DEFINITION
$R(z)$	LP excitation.
$S(z)$	Output of LP vocoder.
G	Gain parameter for LP excitation.
$e(n)$	LP residual excitation.
$A(z)$	LP error transfer function.
a_k	LP filter coefficients.
α_k	LP predictor coefficients.
$x(k)$	Input variable.
$y(k)$	Output variable.
$H(j\omega)$	Variable transfer function.
$H_1(j\omega)$	COMB transfer function.
N	Design parameter of COMB filter.
$w(k)$	Intermediate signal of modified COMB filter.
$z(k)$	Intermediate signal of modified COMB filter.
$H_2(j\omega)$	Intermediate transfer function of modified COMB filter.
$H_3(j\omega)$	Intermediate transfer function of modified COMB filter.

LIST OF SYMBOLS (continued)

SYMBOL	DEFINITION
θ_1	Design parameter of modified COMB filter.
a	Design parameter of all-pole IIR filter.
E(J)	Energy waveform of residual excitation.
H(I)	Lowpass filter transfer function.
a_n	Discrete cosine coefficients.
$H_P(f)$	Spectrum of sample pulse from periodic component.
$H_N(f)$	Spectrum of non-periodic component.
EP,E4	Energy contained in sample pulse.
EN,E5	Energy contained in non-periodic component.

CHAPTER 1

INTRODUCTION

The objective of this investigation was to develop and test improved excitation models for use in speech analysis/synthesis systems (vocoders). One of the biggest drawbacks to widespread use of vocoders is the lack of quality in synthetic speech, which, in turn, is related to the lack of accuracy in commonly used pulse/noise excitations. Of the available vocoders, the Linear Predictive (LP) vocoder is the most commonly used and highly accurate. Thus, an LP vocoder was used as the basic speech analysis/synthesis system for developing improved excitation models. An excitation model was developed which allowed a mixture of pulses and noise, rather than all pulses or all noise for each segment of the excitation. In addition, a method was developed to improve the timing accuracy of the pulses in this excitation model.

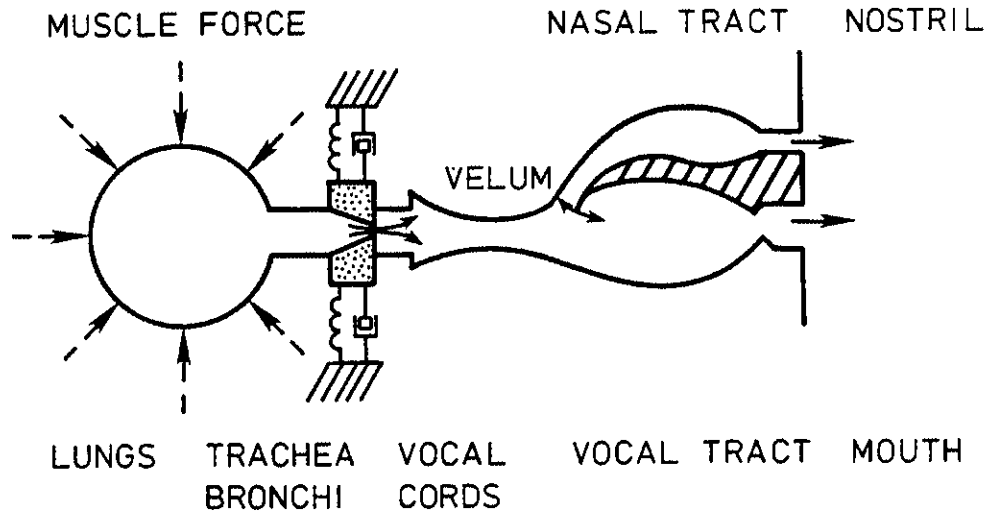
This thesis contains five chapters. The first chapter contains introductory and background materials on Linear Predictive (LP) vocoders. The second chapter discusses the analysis techniques developed and investigated for LP synthesis. The third chapter discusses the models developed and tested. Chapter four describes the psychophysical experiments conducted to compare various voice excitation models and

summarizes the results of these comparisons. It also gives a detailed discussion of the system and software used throughout this research. Finally, Chapter five summarizes the major results and conclusions of this investigation.

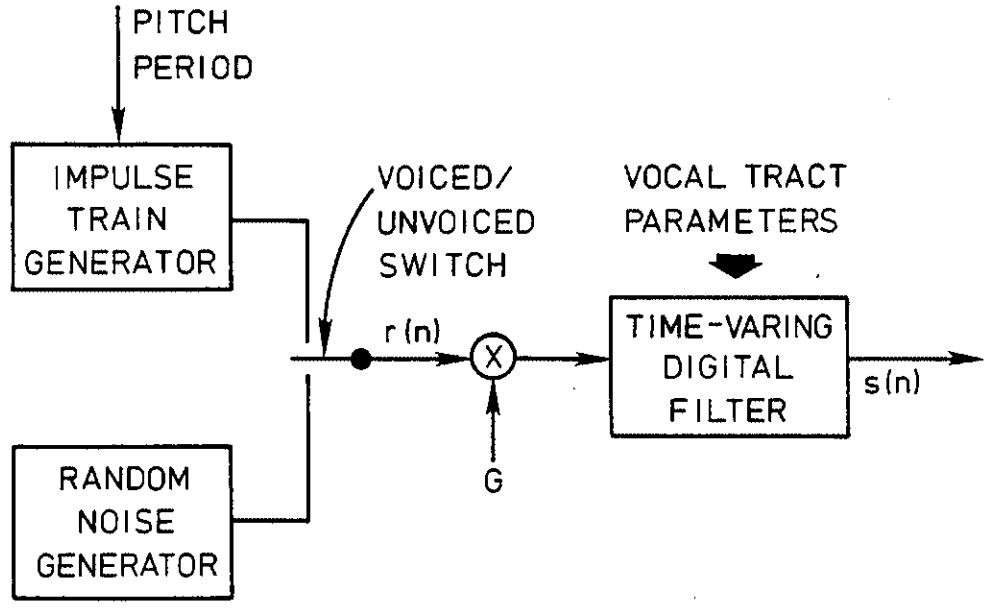
The purpose of the first chapter is to present the theory of speech production and, in particular, the Linear Prediction model of speech production. The Linear Predictive (LP) vocoder was used throughout this thesis for speech analysis and synthesis. Methods for improved speech quality are discussed with respect to improving the LP excitation model. The fourth section describes the processing system used in this investigation. The last section gives the basis upon which this investigation began and its main objectives.

1.1 THE HUMAN SPEECH SYSTEM

Figure 1.1(a) is a representation of the human vocal tract. It shows, in terms of linear system theory, that speech apparatus can be modeled by a source and a system. The source or excitation of air flow is created by the lungs expelling air through the vocal cords. The sound formed from this excitation of air is determined by the system, that is, the resonances of the vocal and nasal cavities. Although speech information is primarily related to the vocal and nasal tract transfer function, speech quality is strongly linked to the excitation signal.



(a)



(b)

Figure 1.1(a) Schematic diagram of human vocal apparatus (after Flanagan, et al., 1970); (b) Discrete-time model of speech production (Schafer, 1972).

Speech sounds are considered to form two major categories, voiced and unvoiced sounds. Voiced sounds occur when air from the lungs is controlled by vocal cord vibrations giving an almost periodic pulse-like excitation. This periodic excitation contains a fundamental frequency known as the pitch frequency, as well as, higher harmonics. Unvoiced sounds occur when air flow from the lungs is passed through a constricted vocal tract causing a noise-like excitation. Thus, the vocal tract excitation is often characterized as purely voiced (periodic) or purely unvoiced (non-periodic). However, actual excitations are often mixed and consist of both types even though one usually dominates. During voiced speech, turbulence in the cavities of the vocal tract cause noise-like effects which become part of the overall excitation.

The vocal tract defines the spectral shaping of the excitation. This spectral shape contains peaks called formants due to the resonant cavities of the mouth and nose. Thus, the vocal tract can be described as a linear time-invariant system over short-time segments (frames). The air expelled by the lungs through the vocal cords is the wide-band excitation to the system and the speech produced is the output. Thus, the spectral envelope of this linear system must approximate that of the original speech. Depending on the speech sound being produced, the excitation is either periodic or noise-like. Since the vocal system remains relatively constant over short intervals of time, about 5 to 20 msec, parameters describing the vocal tract can be computed for each frame of speech to be produced (Rabiner, 1978).

Figure 1.1(b) depicts a basic discrete-time model for speech production. The input parameters are the filter parameters used to describe the vocal tract and thus, determine spectral shape; the voiced/unvoiced decision, used to determine the type of excitation; the pitch frequency, used to restore the fundamental frequency to voiced sections; and the gain factor, used to adjust the overall excitation gain. Each set of these parameters is recomputed for each new frame of speech.

1.2 LINEAR PREDICTIVE VOCODER

A common speech modeling technique is linear prediction (LP). This method affords an accurate representation with easy implementation. Linear Predictive analysis is used to determine model coefficients, a_k, of an all-pole recursive digital filter and the optimum excitation for this filter. The LP model filter represents the spectral shape of the vocal tract. The transfer function relating the input and output of an all-pole discrete-time system is given by:

$$H(z) = \frac{S(z)}{R(z)} = \frac{G}{1 - \sum_{K=1}^{P} a_K z^{-K}}. \qquad (1.1)$$

In the time domain,

$$s(n) = \sum_{k=1}^{P} a_k s(n-k) + G*r(n). \qquad (1.2)$$

Thus, the output speech s(n) is determined by the input excitation r(n) plus a linear combination of the past P values of s(n). For a p-th order linear predictor, using coefficients α_k, the speech output is

$$s'(n) = \sum_{k=1}^{P} \alpha_k s(n-k). \qquad (1.3)$$

The residual signal, e(n), is the error between the actual output of the all-pole filter, s(n), and the output of the linear predictor, s'(n):

$$e(n) = s(n) - s'(n) = s(n) - \sum_{k=1}^{P} \alpha_k s(n-k). \qquad (1.4)$$

Comparing equations 1.2 and 1.3, it can be seen that if $\alpha_k = a_k$, then e(n) = G*r(n).

The z-transform gives the error transfer function, A(z), where

$$A(z) = \frac{E(z)}{S(z)} = 1 - \sum_{k=1}^{P} \alpha_k z^{-k} \qquad (1.5)$$

The residual, e(n), is the excitation and A(z) represents an inverse filter for H(z) which can be used to compute the residual. As shown in equation 1.5, the inverse filter A(z) is given directly by the α_k's, the LP coefficients. This determines e(n) using a finite impulse response (FIR) filter to filter the actual speech signal. The inverse filtering removes variations in the short-time spectral envelope of the speech -- in fact, the LP coefficients can be viewed as those coefficients which cause the spectrum of the residual to be as flat as possible. Thus, the residual signal, e(n), will have a relatively flat spectrum with temporal characteristics closely related to the natural excitation. That is, during voiced sections, there are pulses at a fundamental frequency and, during unvoiced sections, there is a noise-like excitation.

In LP analysis, the predictor coefficients, α_k, are determined such that the mean square error between the actual and the predicted speech is minimized. This process involves first determining the autocorrelation coefficients of each frame of speech (Rabiner, 1978). As discussed above, the optimum excitation for an LP vocoder, the residual, can easily be determined after the LP coefficients are computed. If the residual is used as the excitation, the synthesized speech will be identical to the original speech, and thus of very high quality. However, if low data rate excitations are desired, the residual must be modeled.

Low data rate pulse/noise models of the residual excitation consist of two components where one component represents the periodic (pulse-like) part of the residual and the other component represents the non-periodic (noise-like) component. The most common model of the excitation consists of impulses at the pitch frequency during voiced frames and white noise during unvoiced frames. Therefore, on a short-time basis, the excitation consists either of an impulse train or white noise. No mixture of the two components is allowed for any frame. Figure 1.1(b) illustrates this model.

One error associated with this model is the timing of the impulses. The timing is determined by the starting location of the first voiced frame. An impulse is placed at that starting point and repeated at the pitch frequency for successively voiced frames. This is referred to as "random starting phase", since the actual residual pulse may not occur at the beginning of the frame but at some location after the starting point. In addition, this model does not allow for any mixture of pulses and noise as would be expected from the actual mechanisms of speech production. Examination of the residual also indicates both periodic and non-periodic components, even for voiced speech. Thus, a more complete model for the LP synthesis excitation should contain a pulse/noise mixture, with variable pulse shapes and variable noise spectra. However, the sum of the components should have a flat short-time spectrum, as required by the basic assumptions underlying LP modeling.

1.3 RELATED RESEARCH

Investigations to improve the pulse/noise excitation model of an LP vocoder have been the subject of much research over the past several years. The pulse/noise excitation model has been cited as a major cause for low quality in synthesized speech. One reason is that the model relies on the accuracy of the pitch estimation. In one study, (McGonegal, et al. 1977) several pitch estimation routines were evaluated by subjective evaluations of LPC synthesized speech. Using these pitch routines, synthetic speech was also compared to the natural speech. The final results indicated distinct pitch problems related to various types of speech events. However, natural speech was always preferred above speech synthesized using the most accurate pitch detection method.

Makhoul, et al. (1978), investigated an excitation consisting of a combination of components. The spectrum of the excitation was divided into periodic and aperiodic bands. A frequency, f_c, was determined for lowpass and highpass filters used on impulses and noise, respectively. That is, the periodic component was lowpass and aperiodic component highpass. These components were summed to form the wide-band excitation source. The results showed that models for separate bands of the spectrum improved the naturalness of synthesized speech. Yet, the filters proved to be too rigid to synthesize high quality speech.

Sambur, et al (1978), compares optimum pulse shapes. Parameters for the rise and decay times on the pulse shapes were varied for many speakers. The results indicated a need for definition of pulse shape. However, this rather ad-hoc model has not been able to produce really high quality speech.

A Fourier series representation of the residual was developed through a modified analysis-synthesis procedure as discussed by Atal and David (1979). This method avoids spectral amplitude and phase distortions by developing a way to restore the spectral distortions created by the LP filter. The results indicate a need to restore the amplitude errors over phase errors. The quality improved only when the original amplitude spectrum was maintained.

Schroeder and Atal (1985) and Atal and Remde (1982) discuss methods of excitation modeling using optimum input sequences of pulses. In Atal and Remde's method, the amplitude and locations of a series of pulses are determined for the excitation such that the mean square error between the original and synthesized speech is minimized. No explicit voiced/unvoiced decision is required. However, for voiced sections, the excitation function is primarily periodic whereas, for the unvoiced sections, the pulses are much more random. The results indicate that natural sound is restored to synthesized speech with this excitation sequence. However, this method is computationally difficult, requires the storage of many pulse amplitudes and positions, and does not result in model parameters which easily relate to actual speech production.

All of these references discuss methods to improve Linear Predictive vocoder speech quality by improving the input excitation model. Each strategy developed models closely resembling the natural excitation as seen in the residual. The investigation reported in this thesis used these analyses, algorithms, models, and results to develop a basis for improving the LP vocoder excitation.

1.4 OVERALL SYSTEM DESCRIPTION

In this section, a description of the basic hardware system used for handling the analysis and synthesis is described. Software packages and routines that existed prior to this investigation and which were adapted for use are also described.

The high-frequency pre-emphasized analog speech signals were sampled at a 10 kHz rate by a 12 bit analog-to-digital converter, and stored on hard disk. A PDP-11/24 digital computer controlled the processing and storage of this information using floating point arithmetic. The synthesis performed in this investigation was not real-time and all speech files were stored on hard disk. A block diagram is shown in Figure 1.2(a). Playback for listening to these files was performed by a 12 bit digital-to-analog converter at a 10 kHz rate. The output of the D/A was high-frequency de-emphasized and lowpass filtered by a seven pole elliptical filter. Figure 1.2(b) illustrates this system in block diagram form.

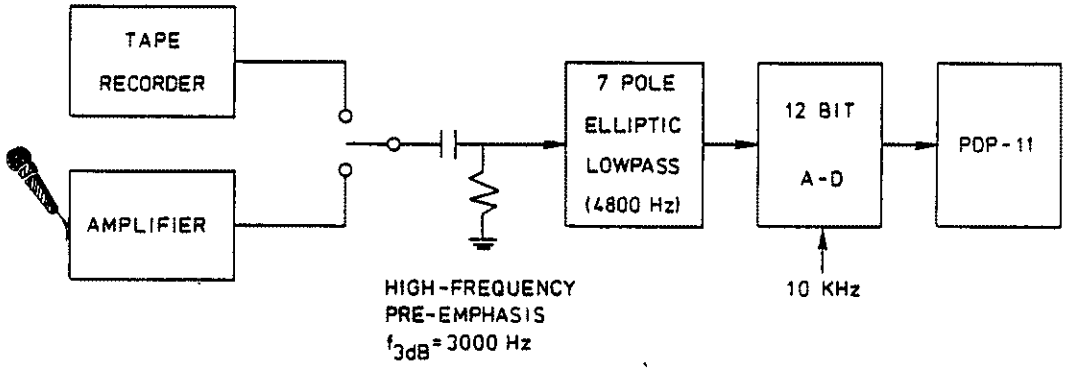


Figure 1.2 (a) Recording system.

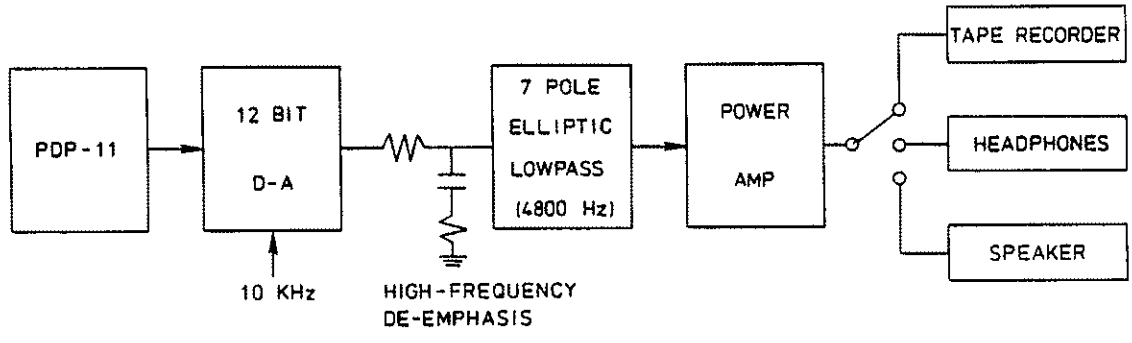


Figure 1.2(b) Playback system.

Software to implement this system was available in the speech laboratory at Old Dominion University. All I/O routines were adapted for use particularly for this investigation. The software for Linear Predictive analysis and synthesis also existed, was adapted and used extensively. The LP analysis routine used the autocorrelation method to calculate the LP coefficients. The LP analysis routine also computed the LP residual and frame energy. Pitch analysis was performed using a form of the SIFT algorithm (Markel, 1972) operating on the LP residual. In this algorithim, the residual is first lowpass filtered at 900Hz, and the normalized autocorrelation coefficients are computed. If these coefficients are below a specified threshold value, the frame is classified as unvoiced. Otherwise, the frame is said to be voiced with a pitch period at the location of the peak value of the autocorrelation. A program was also avialable for smoothing the pitch estimates, in order to eliminate errors made by the pitch routine.

LP synthesis software used the LP coefficients, frame energy, and an excitation signal in order to synthesize speech. The excitation signal could be specified to be either the LP residual, the pulse/noise model using impulses, or the pulse/noise model using the the optimum pulse described by Sambur, et al. (1978). In both analysis and synthesis, up to 14 LP coefficients could be used.

Although the listening experiments determined the final results of this investigation, preliminary experiments involved viewing waveforms to determine software accuracy and to

influence further investigation avenues. A software package called "Speech Lab (SLAB)" was donated by Texas Instruments' Speech Applications Division. SLAB allows the viewing of a disk-stored waveform on a Texas Instruments' Professional Computer (TIPC). It also has options for an instantaneous frequency spectrum, a spectrogram, and for viewing two waveforms simultaneously. All the waveforms and spectrums contained in this thesis were taken directly from the TIPC's screen using SLAB.

A software package call "TTY Communications" was used to link the PDP11 to the TIPC. Using a communications board on the TIPC, a direct line from the PDP11 set up the link between the two computers. Software written for the PDP11 converted the binary formatted speech or excitation files to their ASCII representations and downloaded them to the TIPC, where they were stored on floppy disk. A program written for the TIPC converted the files of ASCII code to a 2's complement binary representation needed for compatibility with SLAB.

1.5 INVESTIGATION STRATEGY

The objective of this investigation was to improve upon the pulse/noise model of excitation for Linear Predictive vocoders. Although this excitation model has the advantage of being extremely simple, it has the disadvantage, as discussed in previous sections, of producing poor quality synthetic speech. In order to test the quality of LP speech obtained using a

pulse/noise excitation model, and the software already available at the Old Dominion University Speech Laboratory, a pilot experiment was conducted. In this experiment, LP vocoders with varying numbers of LP coefficients, corresponding to various degrees of spectral resolution, were implemented using either residual or pulse/noise excitations. Results of a subjective listening test are shown in TABLE 1.1. As can be seen, speech with 8 LP coefficients and residual excitation compared equally in quality to speech synthesized using 14 LP coefficients and pulse/noise excitation. Thus, this experiment also demonstrates the deficiencies in a pulse/noise excitation.

The lack of quality associated with pulse/noise excitation was hypothesized to be due primarily to two effects. The first is the lack of accurate timing. The second was the restriction that each frame be either an impulse train or white noise. We hypothesized, consistent with the study by Makoul, et al. (1978), that a better model would consist of a pulse/noise mixture. The pulses would have varying shapes, but would have their energy concentrated toward lower frequencies. Noise, with energy concentrated at higher frequencies, would be mixed with the pulses to obtain an overall flat excitation spectrum.

In order to test these hypotheses for improving the excitation, methods were developed to separate out the periodic and non-periodic components of the residual excitation. By locating and then defining each of these components, models more complex than impulses or noise and with accurate timing might be determined for better speech quality. Models of both the

TABLE 1.1

Sentence	Preferred A	Preferred B
#1)	0%	100%
A. 14 coefficients, pulse/noise excitation		
B. 12 coefficients, residual excitation		
#2)	0%	100%
A. 14 coefficients, pulse/noise excitation		
B. 10 coefficients, residual excitation		
#3)	44%	56%
A. 14 coefficients, pulse/noise excitation		
B. 8 coefficients, residual excitation		
#4)	75%	25%
A. 14 coefficients, pulse/noise excitation		
B. 6 coefficients, residual excitation		

Table 1.1 Results of a sentence preference experiment comparing speech synthesized with residual excitation versus speech synthesized with pulse/noise excitation. Four speakers and four listeners were used.

periodic and non-periodic components were determined so as to preserve the respective spectral shape of each, but such that the two components would combine for an overall flat frequency spectrum. It was desired that these models be simple to implement with only a small addition to parameter storage. In this study, the effects on speech quality related to timing accuracy in the excitation were also investigated in detail.

CHAPTER 2

EXCITATION ANALYSIS

The main objective of this chapter is to discuss the techniques used for the identification of the periodic and non-periodic excitation components and to present methods for their extraction from the residual signal. Further, potential sources for error associated with these techniques are explored and solved.

2.1 EXCITATION COMPONENTS

The Linear Predictive residual excitation displays several characteristics including a periodic pulse shape during voiced sections with a slowly changing fundamental or pitch frequency, a noise-like component both between pulses and during unvoiced sections, and a relatively flat frequency spectrum. Thus, the residual contains two components: a periodic component and a non-periodic component. From these characteristics, a model for the residual can be determined where the complexity of the model depends on the number of analysis parameters used. The pulse/noise model discussed in Chapter 1 is easy to generate and requires a small number of parameters, but its use in synthesis produces low quality speech.

Determining a more complete model for each of these components is the basis of this investigation. Hopefully, this slightly more complex model would yield better quality speech. Figure 2.1 displays a short segment of speech for two speakers, a female (a,b) and a male (c,d), with their residual excitations. Both segments include unvoiced and voiced frames. These segments represent about 60 milliseconds in time of speech at a sampling rate of 10 kHz. It is easy to observe the periodic pulses and the noise component for each segment of residual.

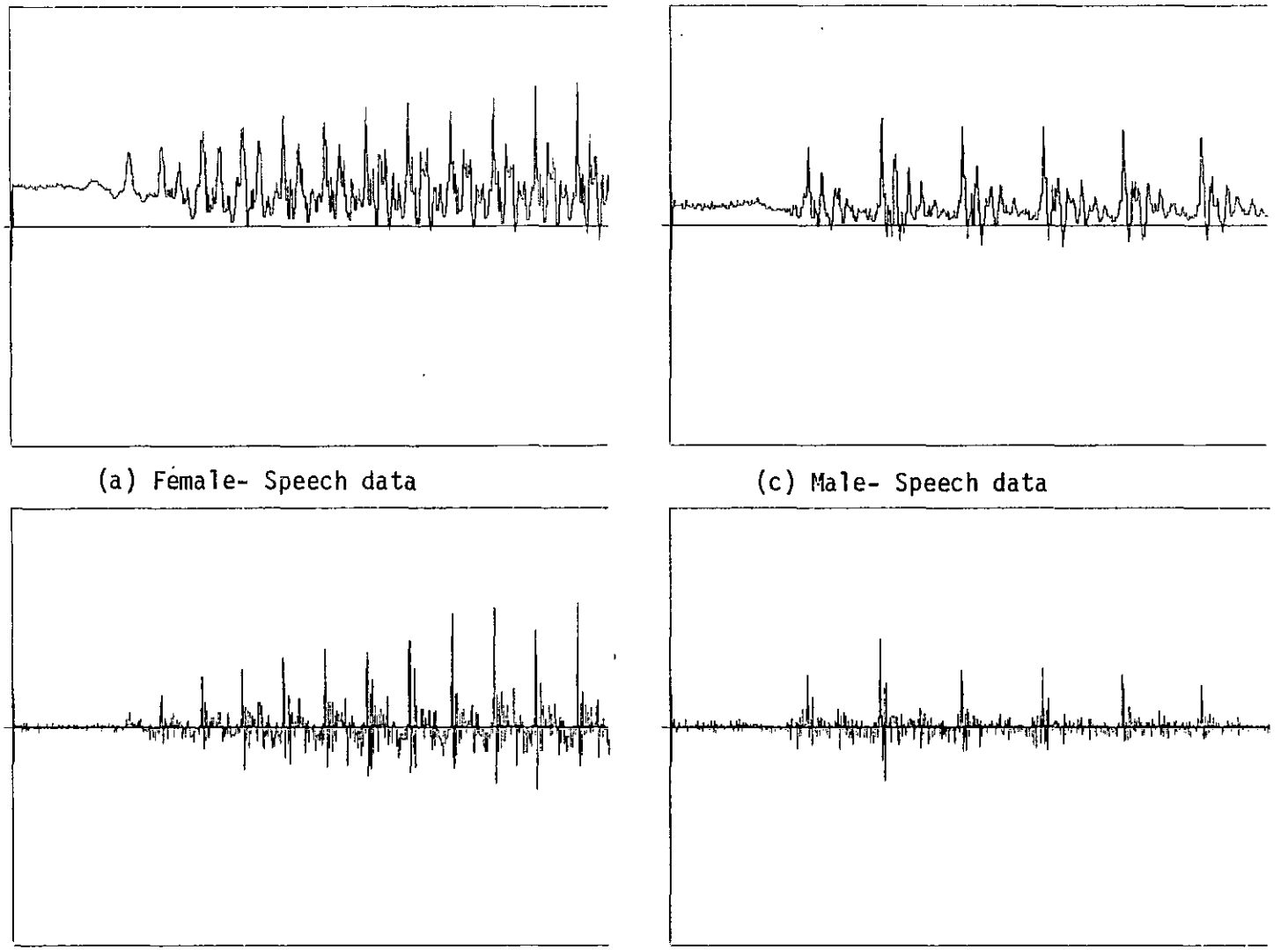
2.2 Extracting Components

The first attempt to separate the periodic component from the non-periodic component of the residual was based on digitally implemented, linear filters. A modified comb filter was used to extract the non-periodic component. A comb filter is an all-zero FIR (Finite Impulse Response) filter which places N zeroes evenly spaced around the unit circle in the z-plane. The value of N is a design parameter. The governing difference equation is

$$y(k) = x(k) - x(k-N). \tag{2.1}$$

Figure 2.2(a) shows the pole/zero diagram and a portion of the magnitude frequency spectrum of a comb filter. The transfer function is given by:

$$H_1(j\omega) = (1 - e^{-j\omega N}). \tag{2.2}$$



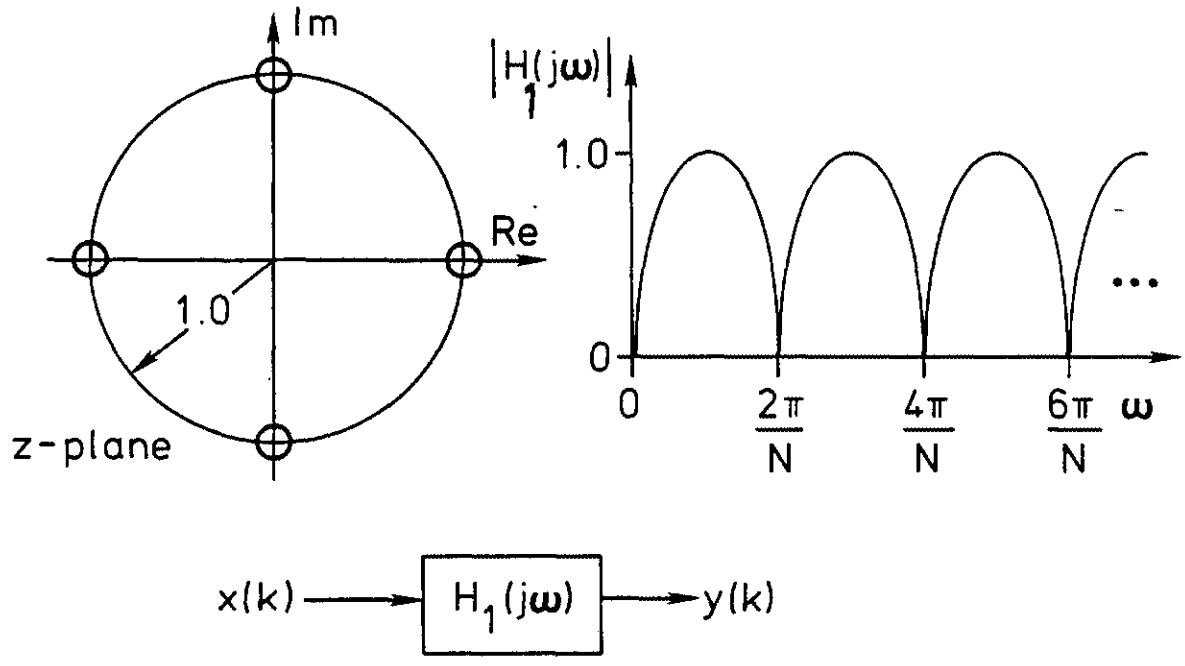
(a) Female- Speech data

(c) Male- Speech data

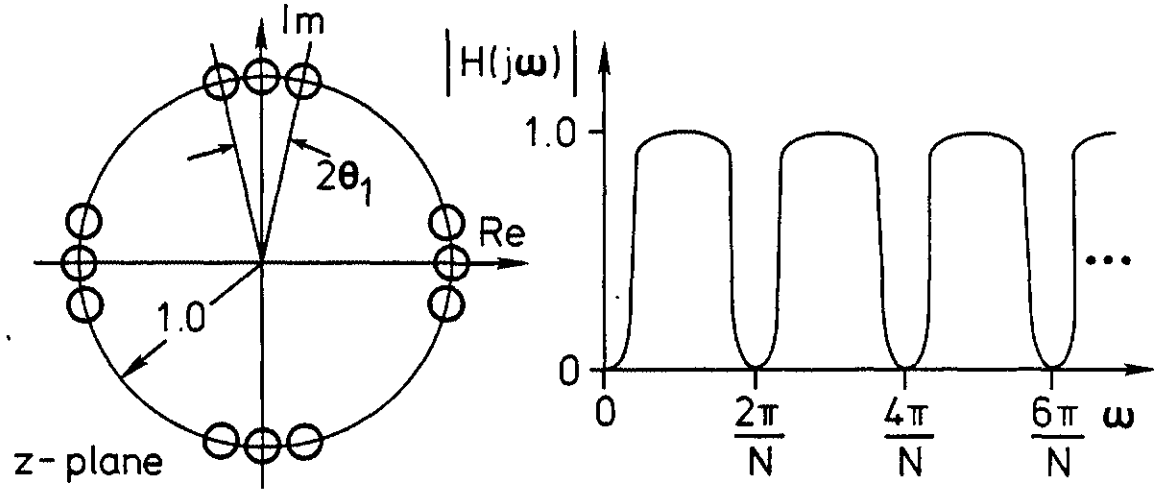
(b) Female- Residual excitation

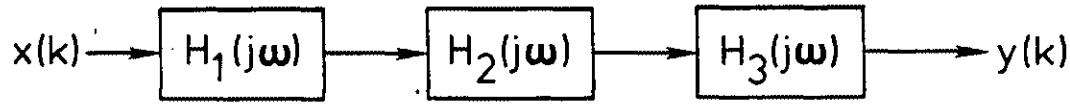
(d) Male- Residual excitation

Figure 2.1 Sample waveforms



(a)





(b)

Figure 2.2 (a) Pole/zero diagram and spectrum of COMB filter.
(b) Pole/zero diagram and spectrum of modified COMB filter.

A comb filter would eliminate the periodic component of a signal of period N, the design parameter. However, the pitch frequency of the residual's periodic component is changing slowly and the value of N may not accurately reflect this period throughout each frame. Thus, the zero points in the frequency spectrum would not necessarily coincide exactly with the pitch frequency. A filter was needed which had a larger zero region in the frequency spectrum, or a "valley" between passbands. Therefore, the comb filter was modified to compensate for small differences between the designed value of N and the actual value of the pitch period in a frame of residual.

The modified comb filter places a group of three zeroes at each of the previous zero locations by using three comb filters in series. The new filter employs a small frequency shift of two zeroes from the zero locations around the unit circle. The first comb filter has no shift and places the first set of zeroes on the unit circle. The second comb shifts another set of zeroes in the positive direction of the unit circle and the third comb filter shifts in the negative direction for symmetry around the first set of zeroes. Figure 2.2 (b) displays the pole/zero diagram and a portion of its magnitude frequency spectrum. The three difference equations which govern this modified comb filter are

$$w(k) = x(k) - x(k-N) \tag{2.3}$$

$$z(k) = w(k) - e^{(-jN\Theta_1)}\, w(k-N) \tag{2.4}$$

$$y(k) = z(k) - e^{(jN\Theta_1)}\, z(k-N), \tag{2.5}$$

where θ_1 and N are design parameters. The second and third equations can be combined to give an equation involving only real variables. As illustrated by figure 2.2 (b), the spectrum now has a small "valley" between peaks and the peaks become wider at the top. The transfer function is given by

$$H(j\omega) = (1 - e^{-jN\omega})(1 - e^{-jN\theta_1} e^{-jN\omega})(1 - e^{jN\theta_1} e^{-jN\omega})$$

$$= H_1(j\omega) * H_2(j\omega) * H_3(j\omega) \qquad (2.6)$$

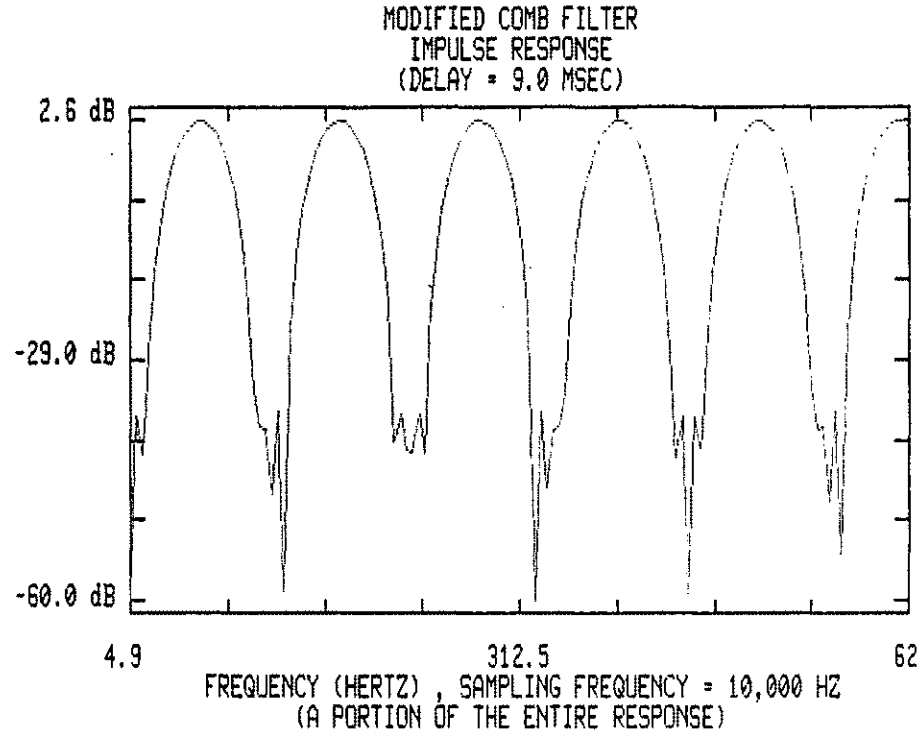
Figure 2.3 (a) depicts the modified comb filter magnitude frequency response for a delay (value of N) of 9.0 msec or 90 samples and a θ_1 value of 0.4 degrees.

Using the small "valleys" between peaks in this modified comb filter to filter out the pulses of the residual, the non-periodic component was found. By subtracting the non-periodic component from the total signal, the result would seem to be the periodic component. However, due to the nature of the phase of this filter, this subtraction does not isolate the periodic component. In order to extract the periodic component, an all-pole IIR filter was implemented in the following manner:

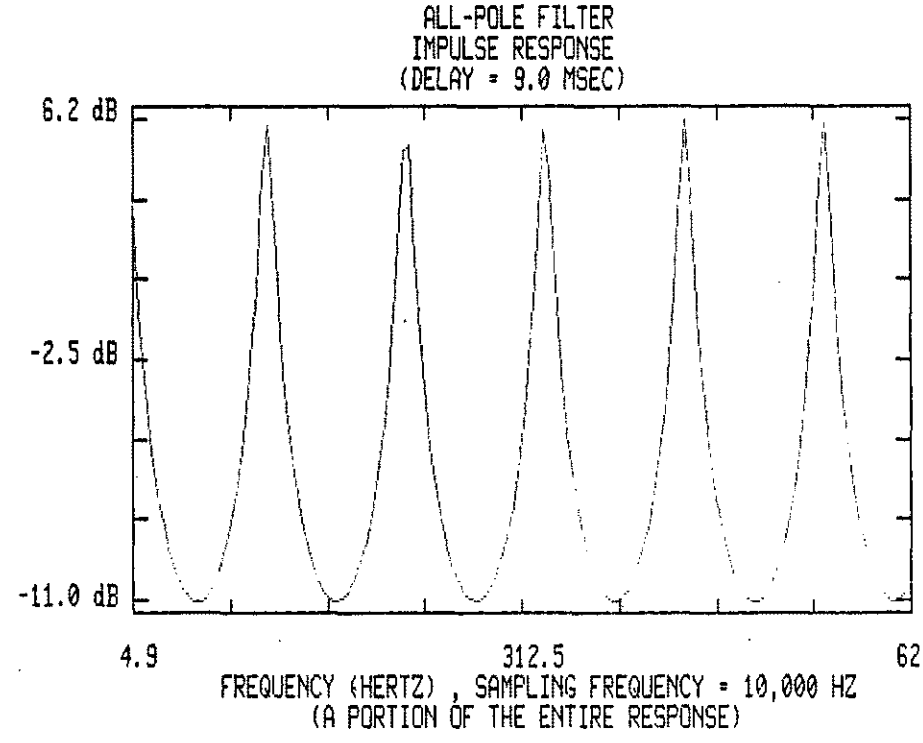
$$H(j\omega) = \frac{e^{-jN\omega}}{e^{-jN\omega} - a^N} = \frac{Y(z)}{X(z)}, \qquad (2.7)$$

corresponding to the difference equation

$$y(k) = x(k) - a^N y(k-N). \qquad (2.8)$$



(a)



(b)

Figure 2.3 (a) Modified COMB filter magnitude frequency response. (b) All-pole filter magnitude frequency response.

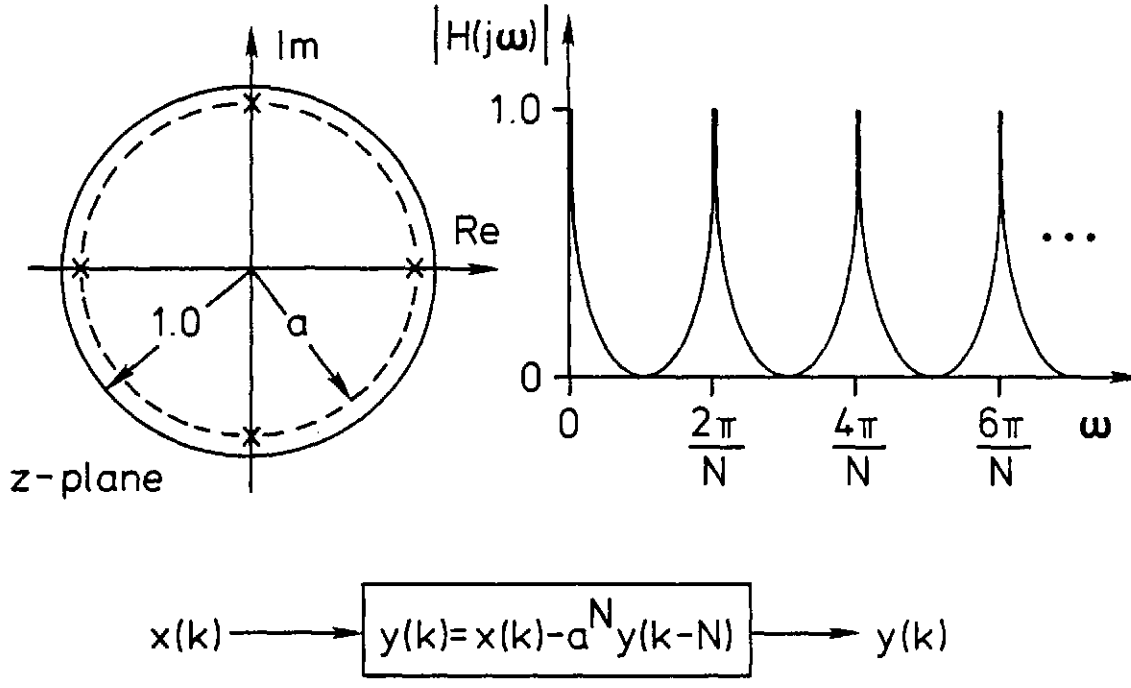
In this case, N poles are placed around the a-radius circle in the z-plane. Figure 2.4 depicts the all-pole filter's pole/zero diagram and a portion of its magnitude frequency spectrum. Figure 2.3 (b) shows the magnitude frequency response for a delay of 9.0 msecs or 90 samples with an 'a' value of 0.997.

The programs COMB and FILT, listed in the Appendix, were used to implement the modified comb filter and the all-pole filter, respectively. The programs continuously filter the residual for successively voiced frames. A buffer of previous input was always stored and used for the next frame's analysis. The inputs to these programs are the ratio of shift in degrees to number of degrees between each of the first set of zeroes for COMB and the the value of the 'a' radius for FILT. The value of N, the delay, is computed from the pitch period found in a previous program. The delay is a number of samples which is derived from the pitch frequency and the sampling rate:

$$N = (\text{Sampling rate}) / (\text{Pitch frequency}) \qquad (2.9)$$

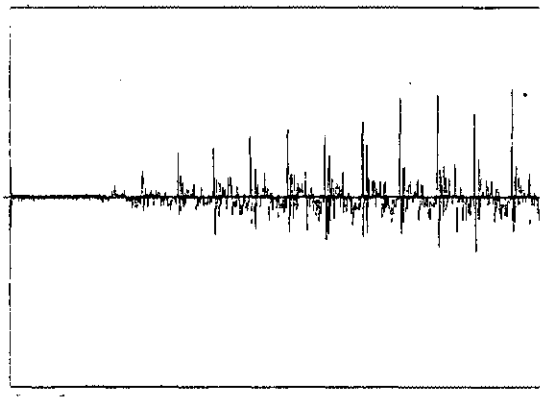
Figures 2.5 and 2.6 illustrate these filters by comparing segments of the residual (a), which is the input of both filters, to the outputs of the modified comb filter (b), and the all-pole filter (c), for a female and a male speaker.

As can be seen in Figures 2.5 and 2.6, the two components were not totally separated. The outputs of the modified comb filter contained some periodicity, while the outputs of the all-pole filter contained some non-periodic components. Also, some preliminary listening experiments showed that there was not

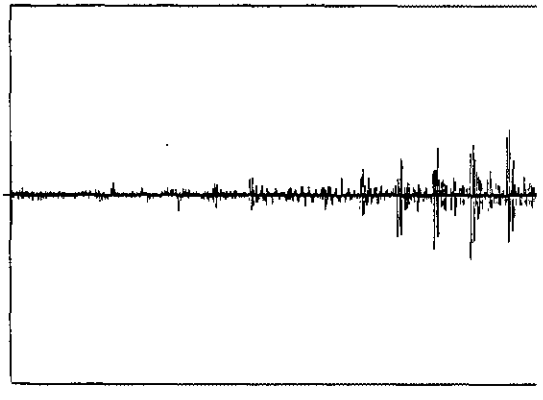


(c)

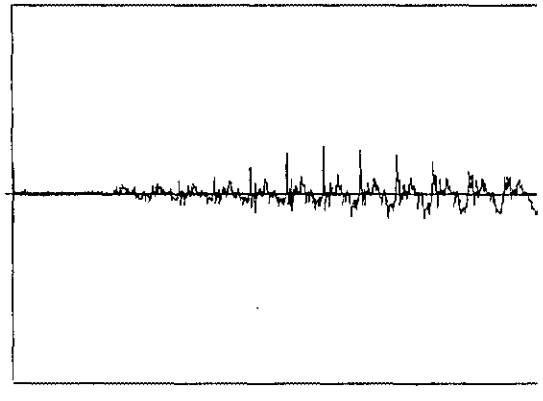
Figure 2.4 Pole/zero diagram and spectrum of all-pole IIR filter.



(a) Residual excitation.

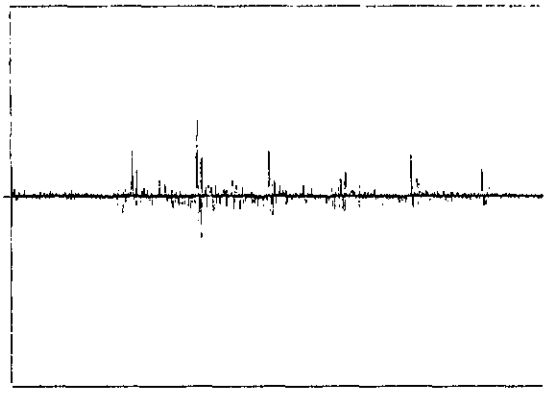


(b) Output of modified COMB.

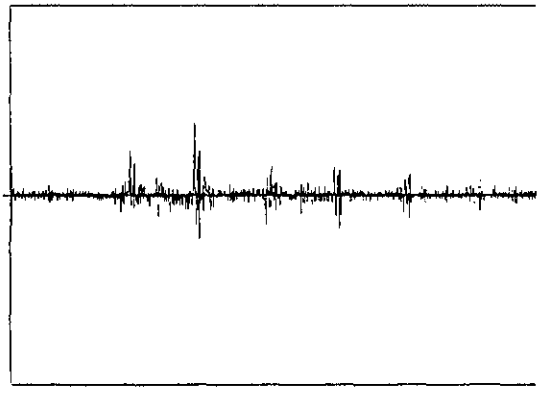


(c) Output of all-pole filter.

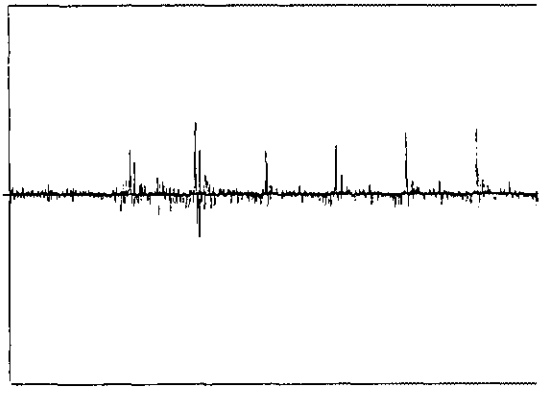
Figure 2.5 Female Speaker.



(a) Residual excitation.



(b) Output of modified COMB.



(c) Output of all-pole filter.

Figure 2.6 Male Speaker.

enough distinct separation. Further, the recombination of the two components produced low quality speech when used as a synthesis excitation.

Both these filters' designs were totally dependent on the correct detection of the fundamental frequency of the periodic component. Small changes in this frequency prevented accurate separation of the periodic and non-periodic components. Further, differing phase characteristics between the two filters made them uncomplementary in the time domain. Therefore, a better algorithm was needed to identify and separate out the periodic pulse-like component. Since this periodic pulse-like component is easily identified in the time-domain, filtering from this aspect can be performed. That is, assuming most of the energy of the periodic component is contained within a short time interval about the large peaks in the residual, and that most of the non-periodic signal is contained in the relatively long intervals between large peaks, the periodic and non-periodic components could be separated provided the exact locations of the large peaks can be determined.

Since the residual contains a pulse-like component, a method was needed to identify each pulse's peak. It was obvious that for each pitch period, the bulk of the energy was contained in the pulse and centered at its peak. Thus, a continuous energy waveform was computed from the residual, by squaring each point of the residual, to identify the peak energy points. These squares were smoothed with a zero-phase, FIR lowpass filter at 0.06 of the sampling frequency to de-emphasize the secondary

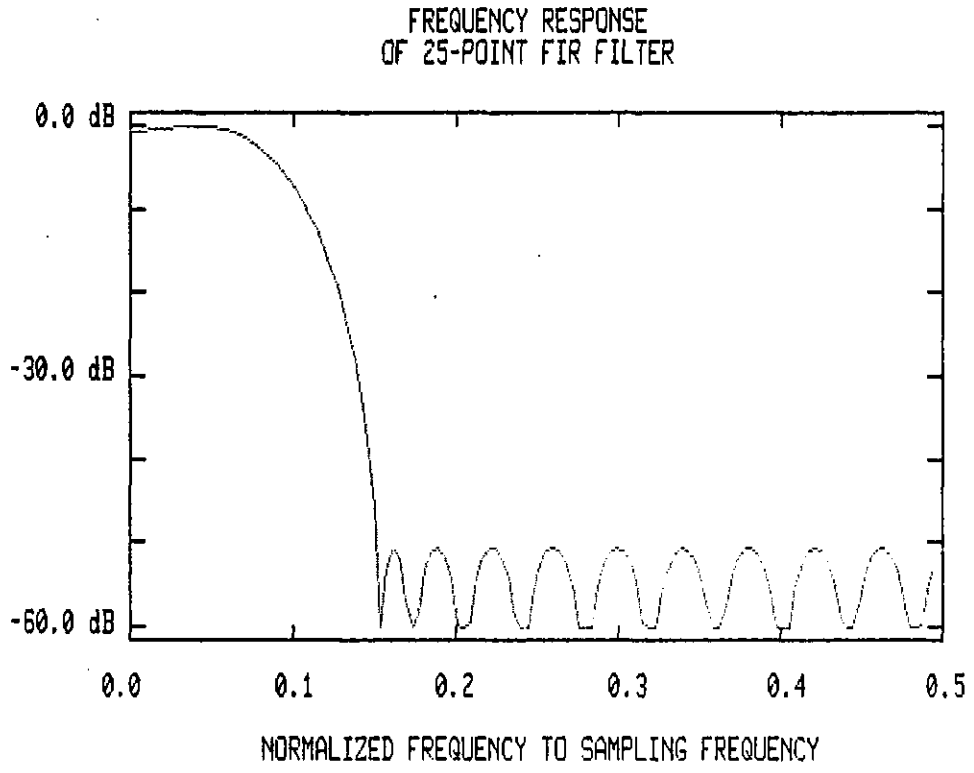
energy peaks so that primary energy peaks could be easily identified. Zero phase shift was possible because of the non-real-time implementation. Thus, the energy peaks aligned temporally with the residual peaks and contained zero shift. The energy value calculated was always located at the same place as the center of the residual points used in its calculation. The zero-phase, low-pass filtered energy waveform was implemented via the following equation:

$$E(J) = \sum_{I=1}^{25} X(I+J-13)^{2} * H(I) \qquad (2.10)$$

where H(*) represents a 25th order FIR lowpass filter, X(*) represents the residual input, and E(*) represents the energy waveform. Figure 2.7 displays the frequency spectrum of this lowpass filter and the filter coefficients and characteristics.

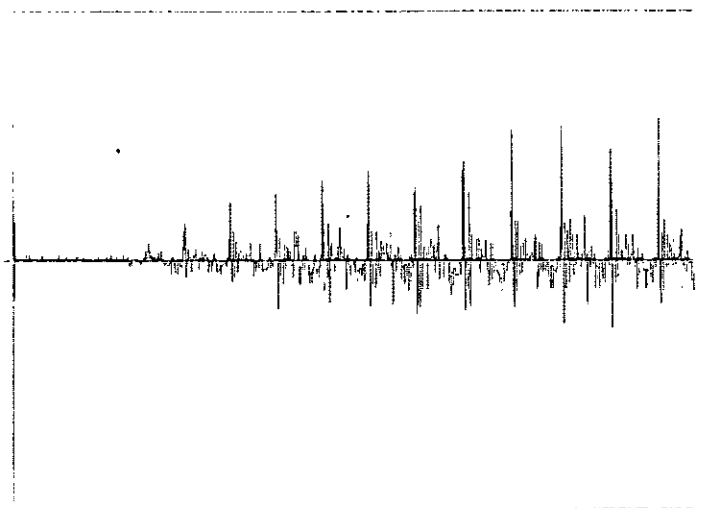
The resulting waveforms are shown in Figure 2.8 with the corresponding input residual excitations for two speakers. The smoothed energy waveform clearly identifies the peaks in the residual excitation. From this waveform, the locations of the peaks were found. Using these locations as perfect timing for any model of the periodic component, a comparison was made between this residual timing and the commonly used timing which had a "random starting phase" in each voiced section.

The program FILTER implemented the calculation of the smoothed energy waveform and selection of the peaks. The peak picking method employed a windowing technique which determined the largest valued point in the selected window of the energy

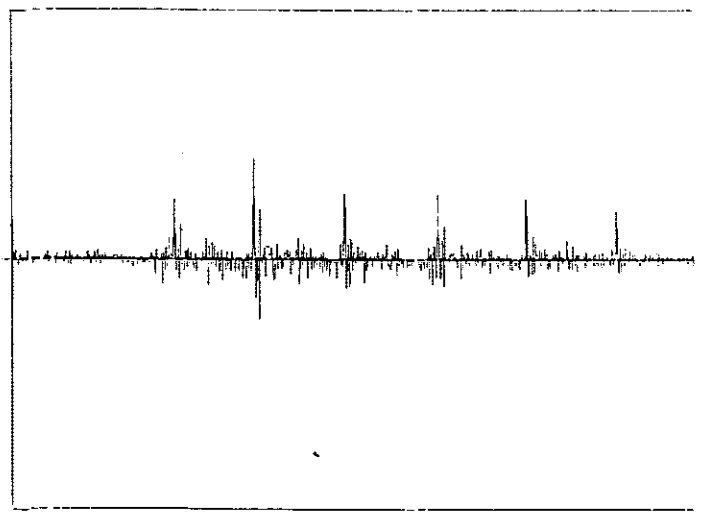



```
**************************************************************************

          FINITE IMPULSE RESPONSE (FIR)
       LINEAR PHASE DIGITAL FILTER DESIGN
          REMEZ EXCHANGE ALGORITHM

                BANDPASS FILTER

            FILTER LENGTH =  25

          ***** IMPULSE RESPONSE *****
        H( 0) =  0.19474268E-02 = H( 24)
        H( 1) = -0.52844197E-03 = H( 23)
        H( 2) = -0.52825962E-02 = H( 22)
        H( 3) = -0.13236498E-01 = H( 21)
        H( 4) = -0.21686191E-01 = H( 20)
        H( 5) = -0.25385415E-01 = H( 19)
        H( 6) = -0.18004123E-01 = H( 18)
        H( 7) =  0.51981132E-02 = H( 17)
        H( 8) =  0.44532232E-01 = H( 16)
        H( 9) =  0.94369218E-01 = H( 15)
        H(10) =  0.14394872E+00 = H( 14)
        H(11) =  0.18058720E+00 = H( 13)
        H(12) =  0.19409601E+00 = H( 12)

                          BAND  1        BAND  2
LOWER BAND EDGE        0.0000000      0.1500000
UPPER BAND EDGE        0.0600000      0.5000000
DESIRED VALUE          1.0000000      0.0000000
WEIGHTING              1.0000002     11.0000000
DEVIATION              0.0329847      0.0029986
DEVIATION IN DB        0.2818781    -50.4616089

EXTREMAL FREQUENCIES--MAXIMA OF THE ERROR CURVE
   0.0000000   0.0408654   0.0600000   0.1500000   0.1596154
   0.1860576   0.2197114   0.2581729   0.2966346   0.3375001
   0.3783656   0.4192312   0.4600967   0.5000000

**************************************************************************
```

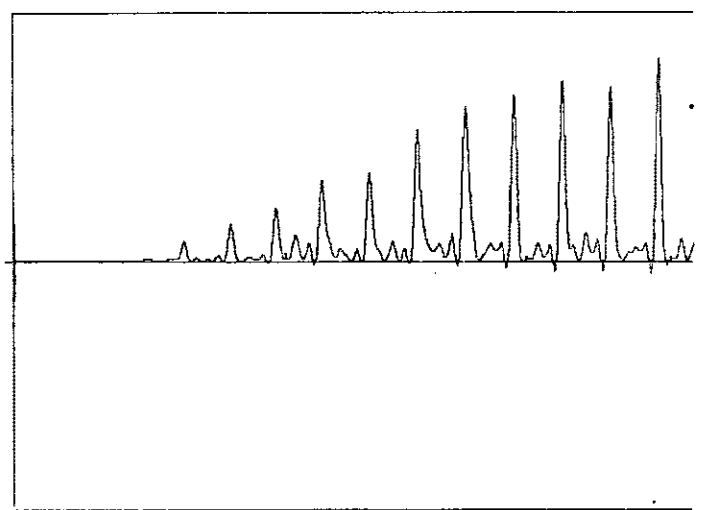
Figure 2.7 Frequency response of 25-point lowpass filter and program design.



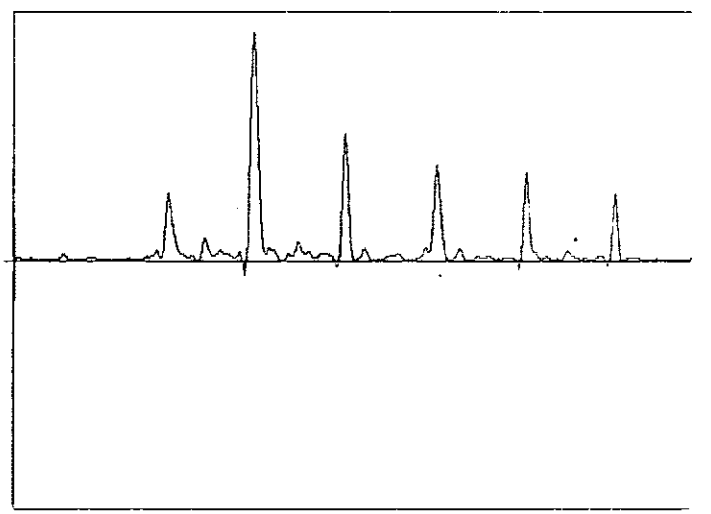
(a) Female- Residual excitation.



(c) Male- Residual excitation.



(b) Female- Energy waveform.



(d) Male- Energy waveform.

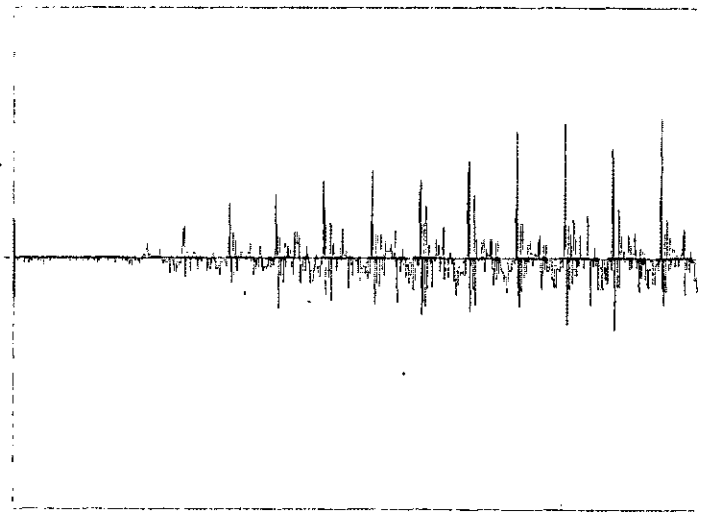
Figure 2.8

waveform. Each window length was a pitch period and the location of each peak was stored. The window is shifted by one pitch period less ten points to overlap with the previous window for exact peak detection. If a pulse should occur on the boundary of one window, the next window would locate a possible larger valued peak. In this case, the CLEAN routine would clean out the smaller valued of the two locations since this was not a true peak. One-point impulses were placed at each location as the pulse models. The operation of this routine was experimentally verified using a software package which allowed the identified pulse locations to be viewed in conjunction with the energy waveform.

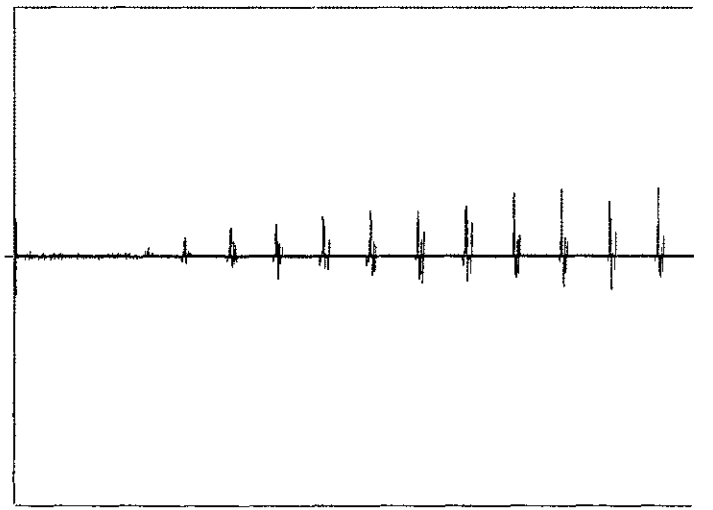
Finally, using the locations of the energy peaks, an N point pulse was selected from the residual for another excitation. Thus, by selecting a number of points from the residual to form the shape of the pulse, the periodic component was extracted. Three values of N were used and tested: twenty points, ten points, and five points. Figures 2.9 and 2.10 show, for a female and a male speaker, respectively, the residual excitations and the new excitations using 20-, 10-, and 5-point pulses.

2.3 ERRORS IN EXTRACTION

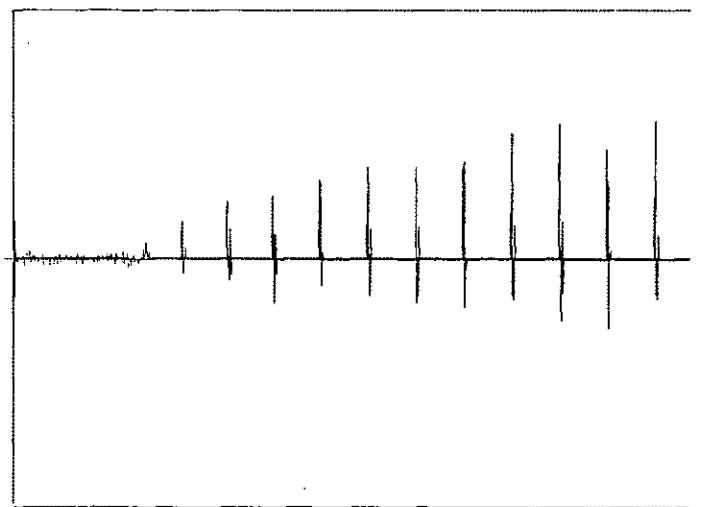
The possibility for errors to occur in these methods has been minimized. However, some inherent errors can occur. One error already mentioned was the lack of complete separation of



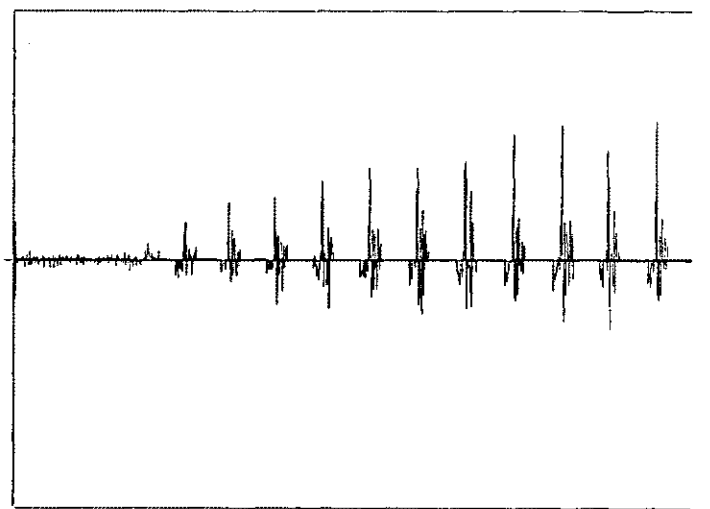
(a) Residual excitation.



(c) Ten point pulses.

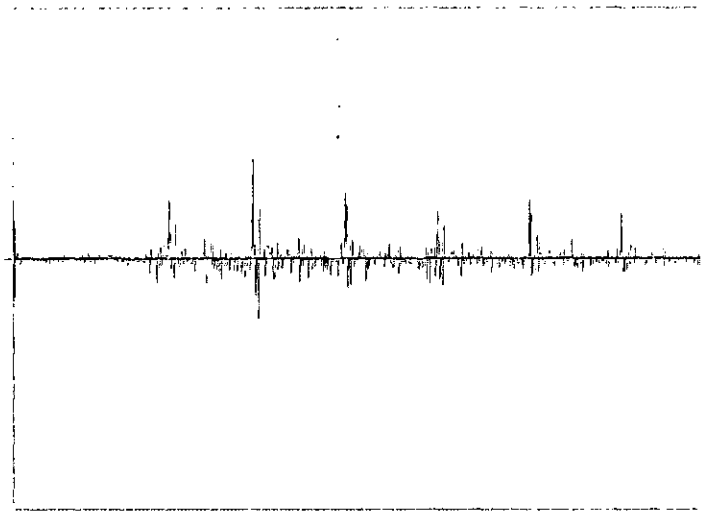


(b) Five point pulses.

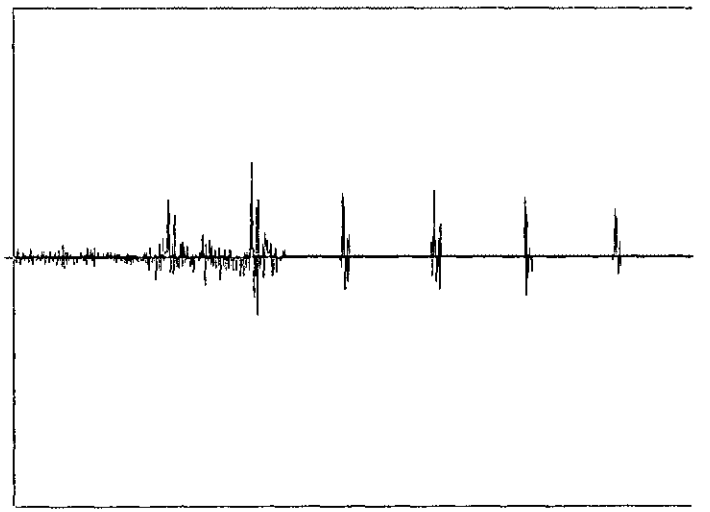


(d) Twenty point pulses.

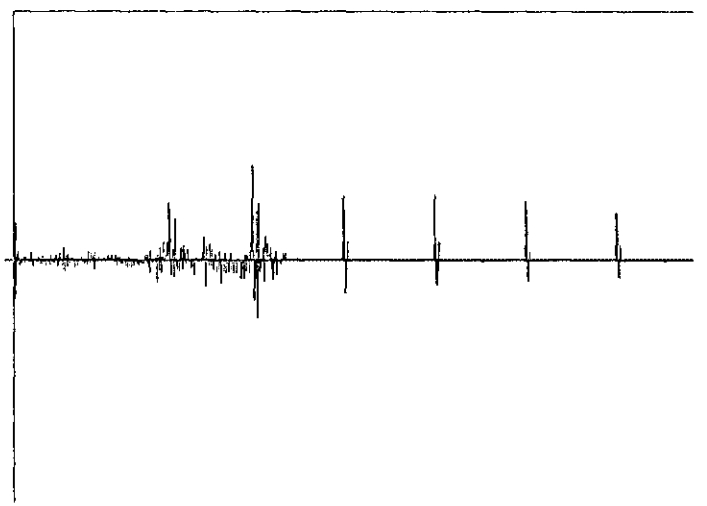
Figure 2.9 Female speaker.



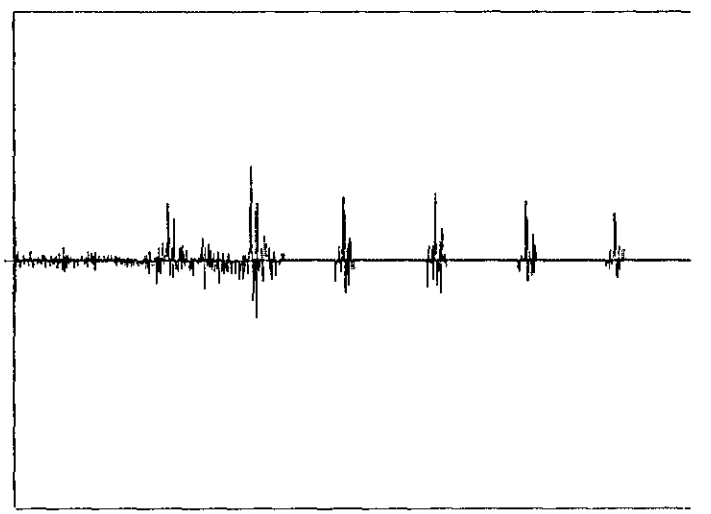
(a) Residual excitation.



(c) Ten point pulses.



(b) Five point pulses.



(d) Twenty point pulses.

Figure 2.10 Male speaker.

the two components using the linear filters. Figure 2.11 illustrates this by displaying the frequency spectrum of the combination of the comb and all-pole filters. This spectrum should be flat at 0 dB to correctly sum the two components. Since this was not the case, it indicates why the separation was not completely successful.

The desire for correct timing of the periodic component is obvious. Figure 2.12 compares the commonly used pulse timing to the precise residual timing found from the peak energy locations. By comparing these excitations to the residual, it is easy to observe the temporal alignment between the impulses using the residual timing and the residual pulses themselves. The impulse excitations shown in Figure 2.12 (c) and (f) do not align themselves with the residual's pulses. These impulses have a temporal shift associated with the "random start phase" as discussed in Chapter 1.

Missing a pulse is another error which degrades speech quality. A "clean-up" routine restores any missing peaks, as well as eliminates extraneous ones. Finally, Figure 2.13 depicts the frequency spectrum of the residual, a twenty-point pulse, a ten-point pulse, and a five-point pulse. As can be seen, the residual's spectrum is relatively flat, yet the five-, ten-, and twenty- point pulses show an errosion of the higher frequencies. The five- and ten- point pulses show much more change of spectrum than the twenty-point pulses from the residual. This indicates that too few number of points per pulse does not define it enough and lacks some necessary higher

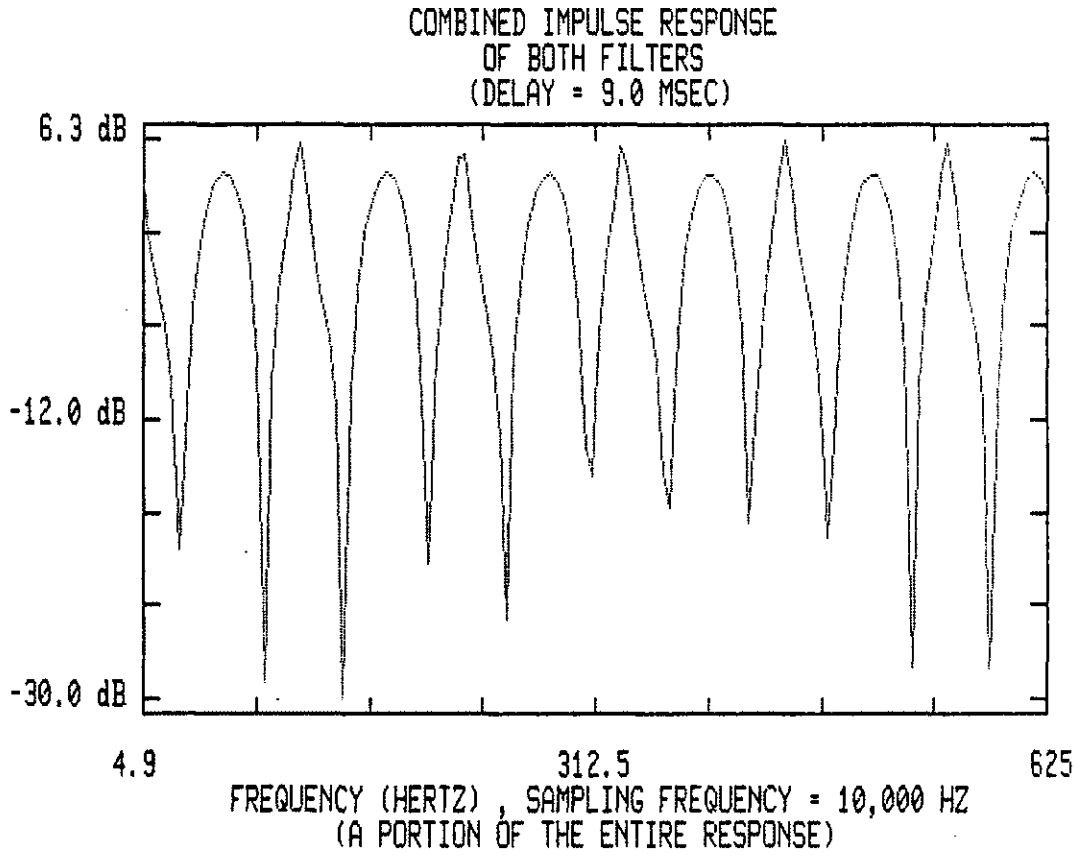
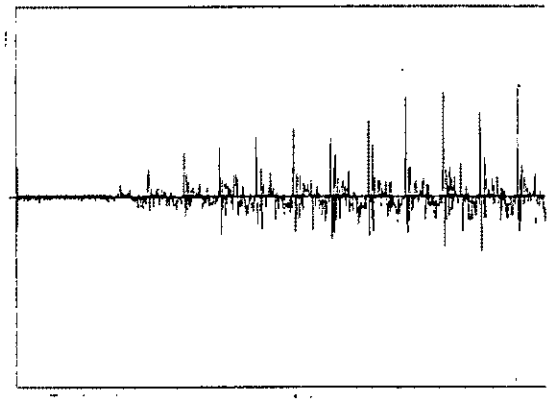
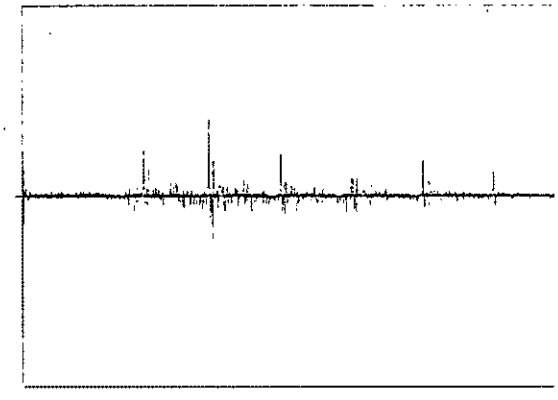


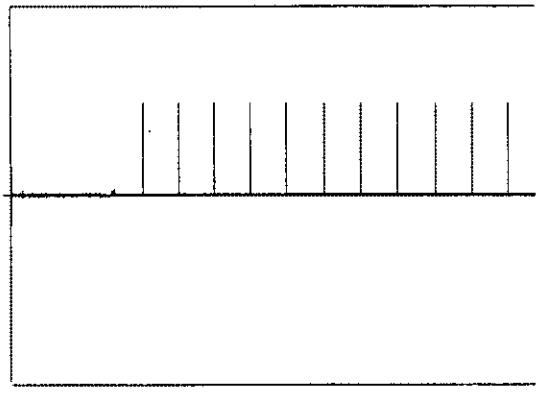
Figure 2.11 Frequency response of the combination of modified COMB and all-pole filters.



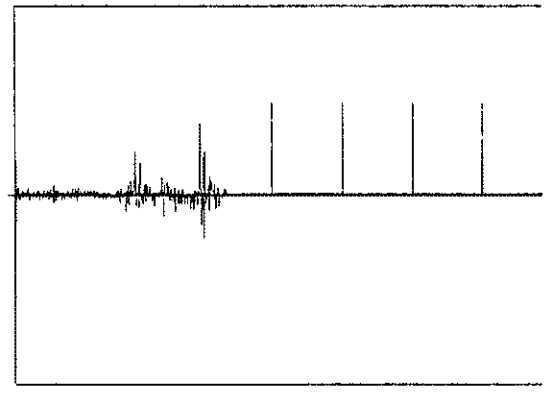
(a) Female- Residual excitation.



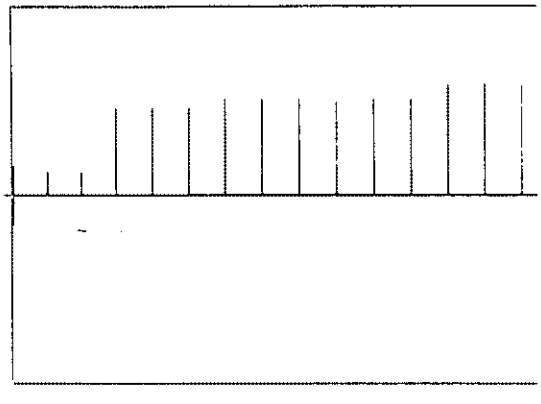
(d) Male- Residual excitation.



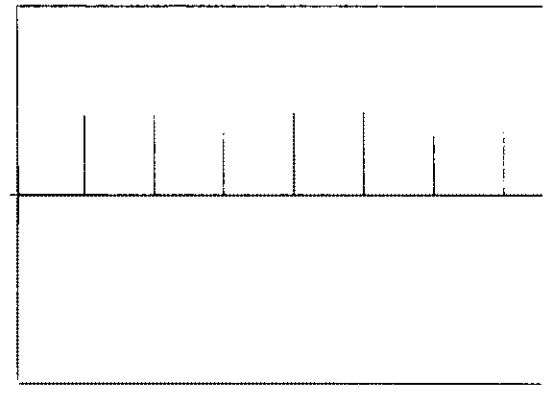
(b) Female- Impulses with timing.



(e) Male- Impulses with timing.

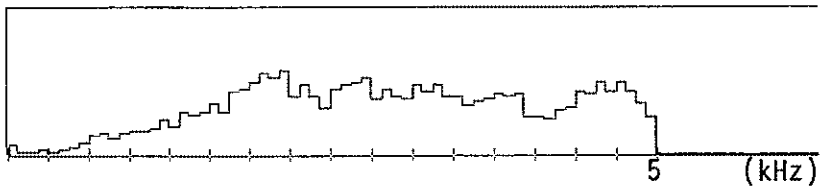


(c) Female- Impulses with no timing.

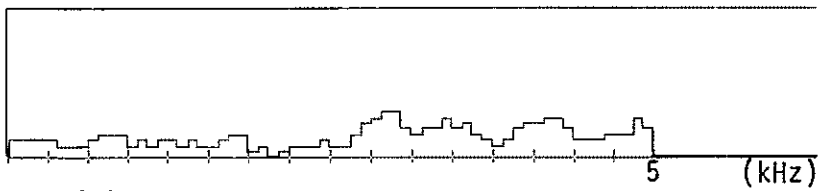


(f) Male- Impulses with no timing.

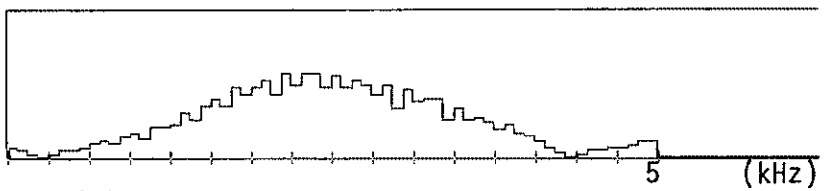
Figure 2.12



(a) Residual frequency spectrum.

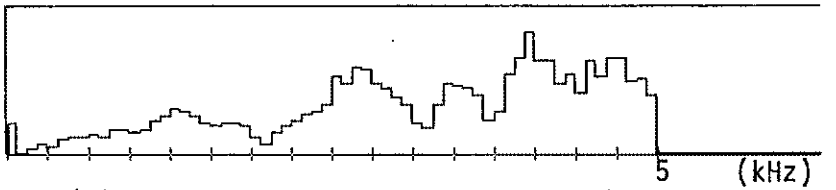


(b) Twenty-point pulse spectrum.

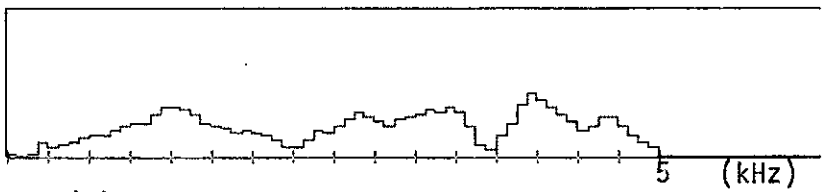


(c) Ten-point pulse spectrum.

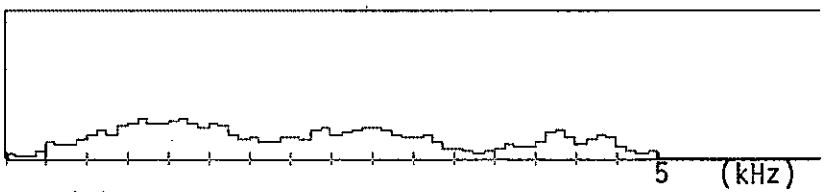
(d) Five-point pulse spectrum.



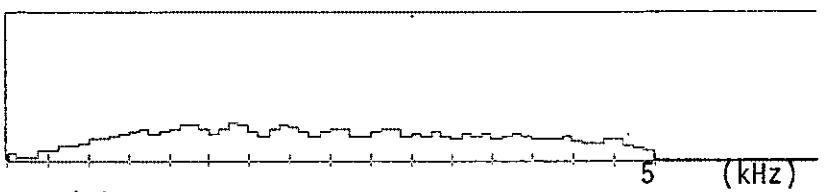
(e) Residual frequency spectrum.



(f) Twenty-point pulse spectrum.



(g) Ten-point pulse spectrum.



(h) Five-point pulse spectrum.

Figure 2.13: Female-(a),(b),(c),(d); Male-(e),(f),(g),(h).

frequencies in the spectrum. The quality of speech synthesized using these excitations reflected this.

CHAPTER 3

EXCITATION MODELING

The primary emphasis of this chapter is the discussion of the modeling techniques developed to describe the Linear Predictive excitation. The last section gives a description of the analysis/synthesis system used throughout this investigation.

3.1 MODELING THE PERIODIC COMPONENT

Once the components of the residual excitation were extracted, the next step was to determine parameters which would closely model each component. Since the periodic component is the easiest to identify during voiced sections of speech, the first attempt was to use only this part of the residual as the excitation. Using the peak-finding routine to identify peaks in the energy waveform, the locations of the pulses were found and thus, the precise timing of the periodic component was found. Initially, one-point pulses were employed as pulse models at these locations. Thus, this excitation is the commonly used pulse/noise model with the important exception that the timing of the periodic component is precisely aligned with the energy peaks in the residual. Further shape to the pulse was obtained

by using a small number of points from the residual itself, centered about the locations of peak energy. Although these are not direct models, they represent the periodic component of the residual excitation with precise timing. At each peak location, five, ten, or twenty point pulses of the residual were used as the excitation model. The rest of the excitation was set to zero. During the unvoiced frames of speech, the residual was used as the excitation. Each excitation was used to synthesize speech for comparisons in quality.

The next step was to model these pulses for further reduction in parameter storage. In the time domain this model should correspond to a finite length (FIR) sequence; in the frequency domain this model should represent the overall spectral shape of the pulse. The frequency domain requirement relates to the desire to be able to add a pulse spectrum (presumably "lowpass") and a noise spectrum (presumably "highpass") to achieve an overall flat spectrum. The time domain requirement stems from the finite duration of the pulses to be modeled. Both of these requirements are met if discrete cosine coefficients (DCC) are used to describe the power-function encoded spectral magnitude of the pulse (Zahorian and Gordy, 1983). The equation used to encode the spectral magnitude is:

$$|H(j\omega)|^{1/3} = \sum_{n=0}^{N-1} a_n \cos(n\omega) \qquad (3.1)$$

where the a_n's are the DCC coefficients.

The a_n's can be used to easily determine an FIR linear phase filter whose magnitude response corresponds to equation 3.1. However, the actual pulses to be modeled appeared to correspond more closely to minimum phase sequences, with maximum energy toward the beginning of the pulse, rather than linear phase sequences, which would be symmetric about the center. Fortunately, as described by Zahorian and Gordy, it was possible to convert the DCC to a minimum phase FIR filter, although with great computational expense.

In the actual modeling process, the sample pulse closest to the center of the frame was chosen. This sample pulse was transformed to the frequency domain using a 32-point FFT of the pulse padded by an appropriate number of zeroes. This magnitude spectrum was nonlinearly scaled to the 1/3 power before computing the DCC. In the synthesizer, these DCC were used to determine an FIR minimum phase pulse whose peak amplitude occurred at the same time as the peak in the residual energy waveform. For a twenty-point pulse model only five DCC were needed.

3.2 MODELING THE NON-PERIODIC COMPONENT

Having located, modeled, and tested the periodic component alone, a model for the non-periodic component was added. Preliminary listening experiments indicated that some portion of the non-periodic component was needed to restore the spectral shape to the excitation. In one modeling scheme, the frame of

residual with the 20-point pulses removed was analyzed to determine its spectrum. From this spectrum, an FIR linear-phase filter was determined. White noise was "high-pass" filtered with this filter in order to complement the spectrum of the 20-point pulses. The excitation was formed by replacing the actual N-point pulses at their exact locations in the excitation. The rest of the excitation was filled with the spectrum-shaped noise.

The next step was to use a model of the pulses with the model of the non-periodic component. A sample 20-point pulse was taken from the middle of each analysis frame and used as the pulse model. Each output frame used its sample pulse at each pulse location. The non-periodic component was computed by the analysis described above. The excitation was generated in the same manner. In both cases, the energies of the periodic and non-periodic components were computed in order to properly scale the model components for synthesis.

In the final modeling scheme, the center 20-point pulse of the analysis frame was used to calculate the spectra of both the components. The spectrum of the pulse was calculated and inverted to determine the complementary spectrum of the noise. That is, the noise spectrum was computed such that the overall spectrum of the noise plus pulse components would be flat. The program NOISE in the appendix was used to calculate the spectrum of the pulse and determine each component's model. The program FIR2 in the appendix generated and filtered the white noise with the spectrum determined in NOISE. This provided the non-periodic component.

The following method was used to calculate the inverted spectrum of the noise given the pulse spectrum. The energies of the periodic component, the non-periodic component, and the entire frame of excitation, respectively, are calculated:

$$EP = \sum_{n=0}^{N1-1} x_p^2(n); \tag{3.1}$$

$$EN = \sum_{k=0}^{N2-1} x_n^2(k); \tag{3.2}$$

and

$$E = \sum_{j=0}^{N-1} x^2(j); \tag{3.3}$$

where the value of N is the length of the analysis frame. The value of N1 is the number of points that made up the pulses in this frame; the value of N2 is the number of points in the frame minus N1.

The frequency spectrum of the pulse also provided its energy, defined by

$$\sum_{i=0}^{L-1} H_p^2(f_i) = E4. \tag{3.4}$$

where the value of L is the length of the FFT used. Before representing the noise spectrum by discrete cosine coefficients, the approximate level of the noise spectrum was found as follows. First, the complementary spectrum's energy was set equal to half of the energy of the periodic component times the ratio of the non-periodic to periodic energy. This half ratio was determined empirically from preliminary listening experiments. Because some of the periodic component's energy is contained in the portion not selected with the pulse, addition of all of the noise energy would add more non-periodic components than necessary. Figure 3.1 displays the desired effect. The equation is given below:

$$\sum_{i=0}^{L-1} [K - H_p(f_i)]^2 = 0.5*E4*(EN/EP). \qquad (3.5)$$

Expanding the equation results in

$$L*K^2 - 2K*\sum_{i=0}^{L-1} H_p(f_i) + E4 = 0.5*E4*(EN/EP). \qquad (3.6)$$

Solving this quadratic equation for K gives the value upon which the noise component is based. The equation follows:

$$K^2 - 2/L*\sum_{i=0}^{L-1} H_p(f_i) * K + E4/L*(1-0.5*EN/EP) = 0 \qquad (3.7)$$

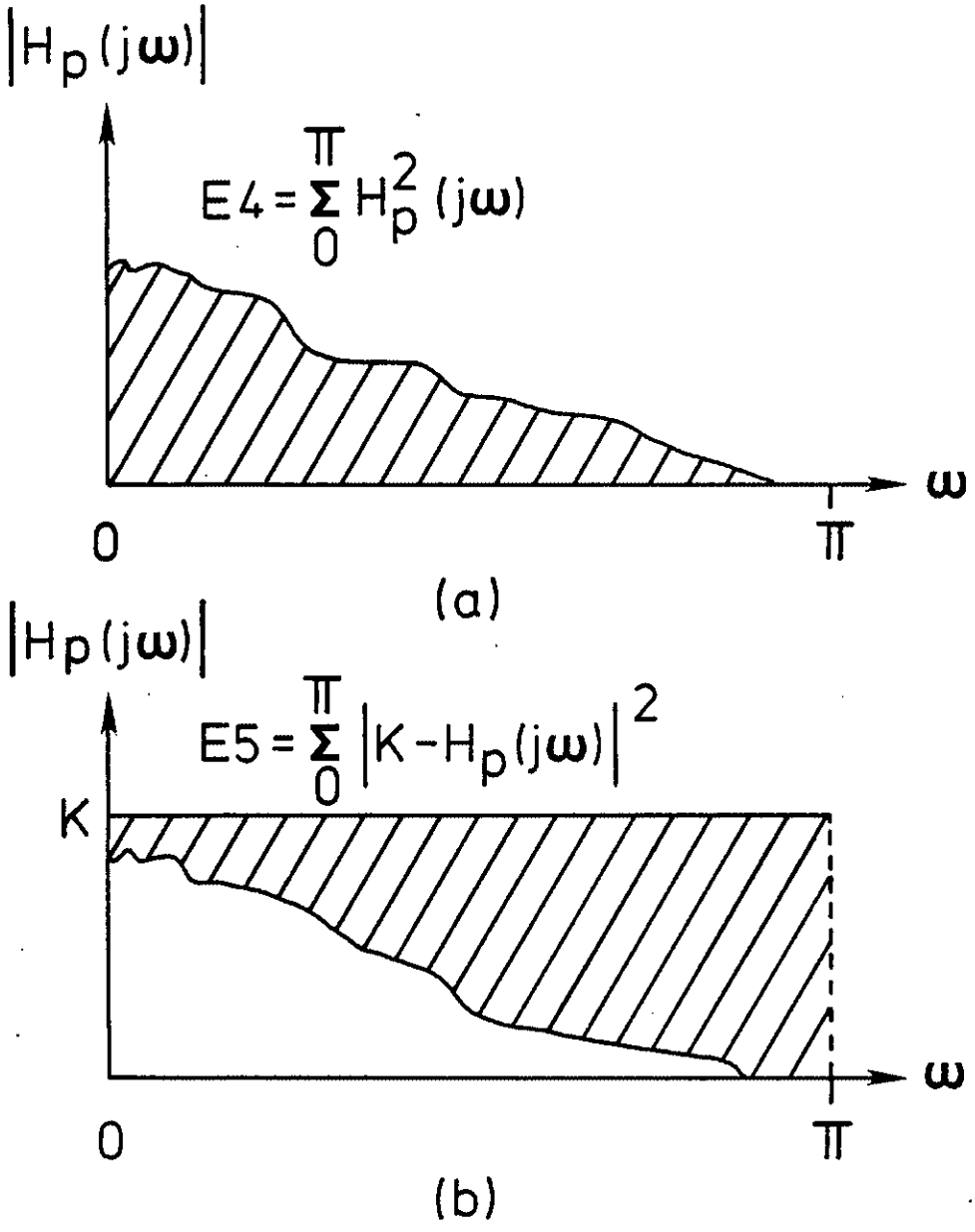


Figure 3.1 (a)Frequency spectrum of pulse and energy equation; (b)Complementary spectrum and energy equation.

The result is

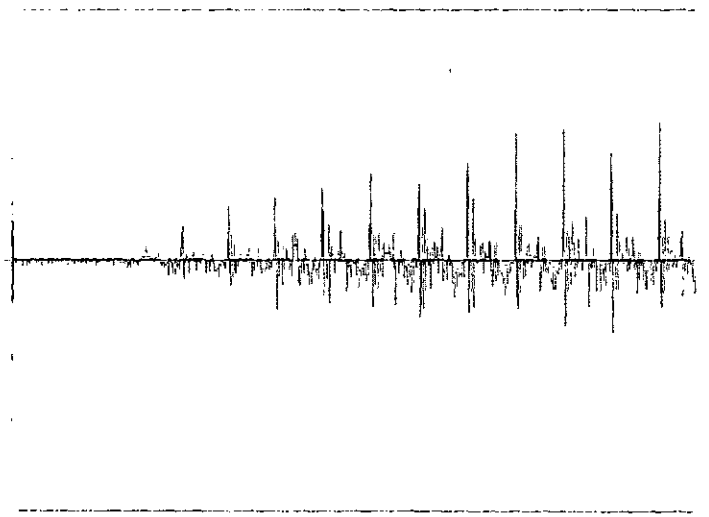
$$K = \sum_{i=0}^{L-1} H_p(f_i)/L + \left(\sum_{i=0}^{L-1} H_p^2(f_i) \right)^2 /L - E4/L*(1-0.5*EN/EP) \quad (3.8)$$

Using this value of K, the spectrum for the added noise can be found

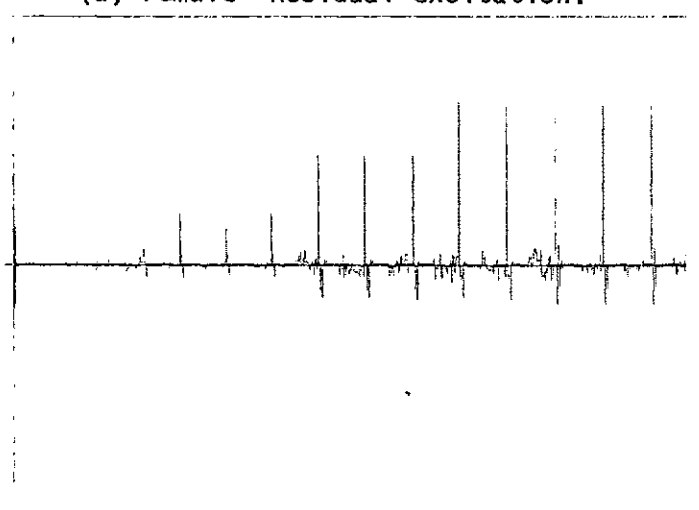
$$H_N(f_i) = K - H_P(f_i). \quad (3.9)$$

This spectrum was nonlinearly scaled to the 1/3 power and modeled by five discrete cosine coefficients.

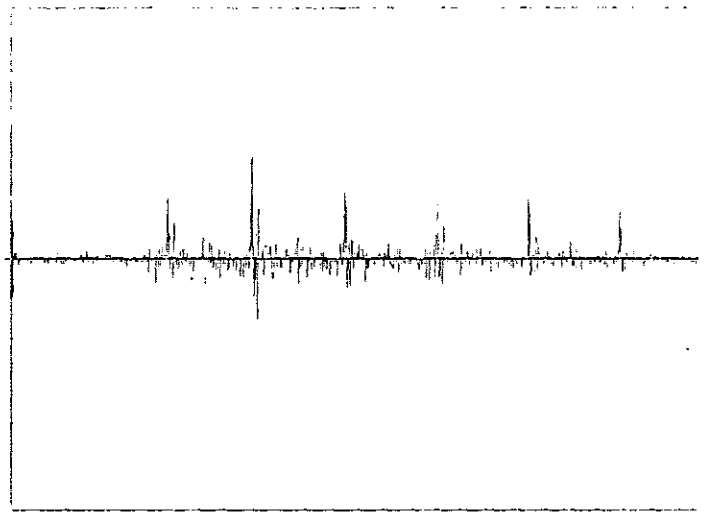
Figure 3.2 shows the residual and the final excitation model containing 20-point modeled pulses plus added spectrum-shaped noise. Figure 3.3 is a block diagram of the system used for the final synthesis model. The excitation used five coefficients for the minimum phase filter to model the pulses in each frame and five coefficients for the linear-phase filter to model the spectrum of the white noise. Other parameters include the pitch, voiced/unvoiced decision, the respective gains of the periodic and non-periodic components, and the peak locations for the pulse models in each frame for implementing residual timing. Using the excitation created by these parameters, fourteen predictor coefficients formed the LP filter for the final synthesis.



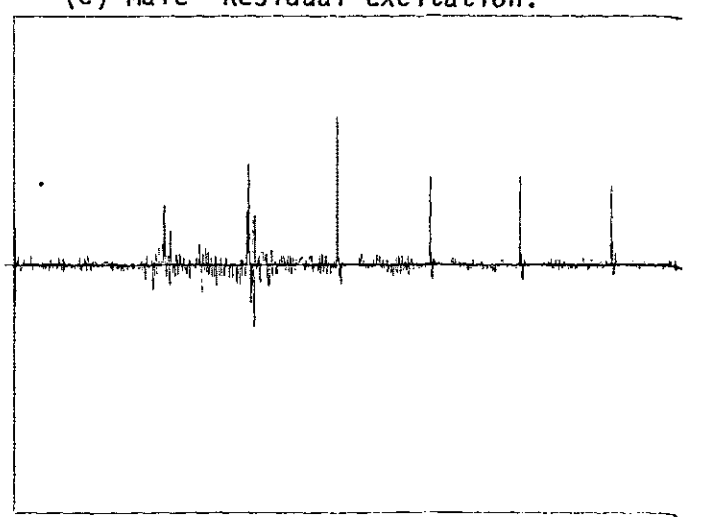
(a) Female- Residual excitation.



(b) Twenty point pulse models plus noise.



(c) Male- Residual excitation.



(d) Twenty point pulse models plus noise.

Figure 3.2

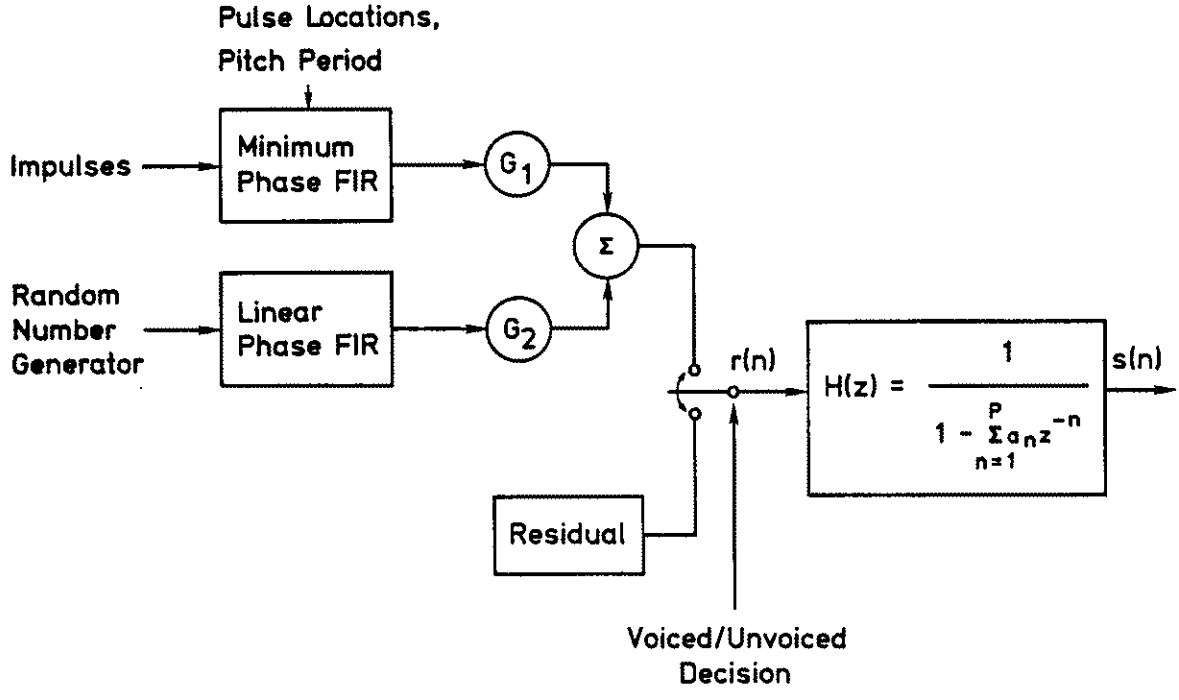


Figure 3.3 Block diagram of speech synthesis using excitation modeling scheme and LP filter.

3.3 ERRORS DUE TO MODELING

The primary errors in the linear filtering approach are related to the far-from-perfect separation of the periodic and non-periodic components and to the combined impulse response of the two filters. Nevertheless, the all-pole, linear filter was used to "extract" the periodic component. This component was experimentally tested as a possible excitation model. However, since the all-pole filter was not completely successful in separating out the periodic component, this excitation was not appropriate for further analysis.

Another possible source of error is the voiced/unvoiced decision. Since the analysis is performed on voiced frames only, an incorrect decision (voiced for unvoiced) would result in a periodic model for an unperiodic frame. This occurs mostly at the beginning and the end of voiced sections of speech.

Another similar error can occur with the pitch estimation. Although pitch usually changes slowly along frames, sometimes the routine to estimate the pitch is incorrect when there is a rapid change in pitch. The peak finding routine depends on the accuracy of the pitch estimation. This error had even more effect on the linear filters, since their design was entirely dependent on the value of the pitch estimation.

When modeling the pulses, a model is developed for one sample pulse in the frame and is repeated at each pulse location. Thus, for a changing pulse shape and size, such as that which occurs at the beginning of a voiced section, this

model does not reflect the true periodic component. In addition, the assumption was made that the phase of the spectrum of the pulse was relatively unimportant. The minimum phase characteristics used for the model pulse was an approximation.

Another possible source of error associated with the time-domain method for separating the periodic and non-periodic components is the assumption that the periodic and non-periodic components of the residual are separated in time. That is, this method assumed that the periodic component was totally contained within a very short time span centered about the peak energy points in the residual while the non-periodic component is totally contained in the intervals between these points. Although this assumption is not totally valid, at least the two components separated in this manner can be summed to give back the original signal.

3.4 SPEECH ANALYSIS/SYNTHESIS SYSTEM

All speech data, as well as excitations, were stored on the PDP-11 digital computer in 155 blocks of 256 two-byte integers. This amount of memory represents about four seconds of speech for a sampling rate of 10 kHz. Each data file was a complete sentence spoken by one person. When analysis on the residual was done, the data was read from memory by analysis frame lengths. For male speakers, the analysis length is 400 points; for females, it is 325 points. Each point is represented by a two-byte integer. Once the analysis is completed on an analysis

frame, only the center 128 points of corresponding output are saved. The next analysis frame begins at the 128th point of the previous analysis frame. Thus, the frame spacing is 128 points, which allows a different 128 points to be written out each time. The output is stored in the same format as the input. The use of 256-point Hamming Windows emphasizes the middle 128 points of an analysis frame. This type of filtering is often used on analysis frames.

Synthesis in every case used a 14th order Linear Predictive filter to achieve the best quality speech possible. This way, it was easier to recognize speech degradation from different excitations and not due to the LP filter. The 14 LP coefficients were determined by an autocorrelation analysis on a 256-point frame of speech weighted by a 256-point Hamming Window.

The pitch period was determined with the same analysis frame lengths as for the residual analysis. The voiced/unvoiced decision follows the calculation of the pitch period by using a normalized autocorrelation of the lowpass filtered residual. If the absolute peak was above 0.45, the frame was judged voiced. This limit was raised to 0.50 for frame energies below 100,000. It was lowered to 0.30 for successive frames whose pitch varied smoothly.

Once the parameters of the LP filter were found, the analysis of the residual to determine an excitation model was made. Given these model parameters and the LP filter

coefficients, an excitation was produced and the speech was synthesized.

CHAPTER 4

EXPERIMENTATION

This chapter describes the experiments performed to evaluate the various excitation models developed in this study. Results of each experiment are tabulated and a brief discussion of results is included. A more general discussion of experimental results is deferred to Chapter 5. Also, the software used to produce the speech segments used in the experiments is described.

4.1 ANALYSIS-SYNTHESIS SOFTWARE

Each speech file was synthesized in three steps. The first step generated the residual excitation in one file and the synthesis parameters in another file. The second step took the residual, performed the desired analysis, and modeled the excitation. The third step used the modeled excitation and the synthesis parameters to synthesize the speech.

In the first step, the routine AUTO calculated the Linear Predictive (LP) coefficients by computing the autocorrelation coefficients from the Hamming-windowed data segment. These coefficients were used to find the LP coefficients. The routine LPCSYN computed the residual using the LP coefficients for a

matched filter and the data segment. The frame energy was also calculated and placed in an output file with the LP coefficients for each 128-point frame. The residual was placed in another output file for analysis. Finally, the pitch period and voiced/unvoiced decision were determined in the routine PEAK7 using the residual as the input signal. A normalized autocorrelation of the 900 Hz lowpass filtered residual was computed to accurately measure these parameters. These analysis programs existed prior to this investigation and were adapted for use here. Each is listed in the appendix.

The second step involved using the residual file to generate a model for the vocoder excitation. An analysis length frame of residual was passed to the routine FILTER which computed the lowpass filtered, running energy waveform. A model for the periodic component was generated using the pulses centered at peak locations. The routine NOISE calculated the necessary noise spectrum and generated the added noise component. The two components were scaled and combined into an output excitation file.

When the linear filters were used initially, the routines COMB and FILT replaced FILTER and NOISE in the program. These filters employed the difference equations discussed in Chapter 2. For all analysis routines, a frame of residual was analyzed and modeled only during voiced frames of speech. During unvoiced frames of speech the residual was written directly into the excitation file.

The final step took the excitation file and the file containing synthesis parameters to implement the LP filter and synthesize speech. The routine LPRES scaled the excitation and called LPCSYN to synthesize speech with the LP filter. Finally, the speech was placed in an output file on the computer. The synthesis software was in existence prior to this investigation and was adapted for use in the experimentation.

4.2 SYNTHESIS EXPERIMENTS

Four experiments were conducted. Three of these consisted of quality comparisons of the same sentence synthesized using two different LP filter excitations. Each comparison was made for four different speakers and played in both orders of comparison. In the fourth experiment, listeners were asked to rank in order of preference four sentences (the same speaker) corresponding to different excitations. In every case, the synthesis was the same using a 14th order LP filter; it was the excitations that were different. Five listeners were used in each experiment.

The experiments in which excitation models were compared offered a perceptual measure of quality of the synthesized speech. The following sentences used in the experiments were chosen for their "balanced" phonetic content and have been used in other speech experiments. (Zahorian, 1978).

1. "Every salt breeze comes from the sea." Male Speaker (PG)

2. "We were away a year ago." Female Speaker (EE)

3. "I was stunned by the beauty of the view." Male Speaker (SZ)

4. "The trouble with swimming is that you can drown." Female Speaker (AG)

4.2(a) Experiment #1

In the first experiment, several basic excitation analysis methods were compared. The two linear filters were judged for their separation of the two components. The peak finding routine determined residual pulse timing for comparison to "random start phase" pulse timing. Also, LP synthesis was compared between speech synthesized using a 14th order filter to determine the pitch period and using a 10th order filter to determine the pitch period.

There were six comparisons in the first experiment of sentences synthesized from several different excitations. The residual (RESID) was compared to the periodic component found from the all-pole filter (PER). This periodic excitation (PER) was then compared to the combination of components found from the modified comb and the all-pole filters (PER/NON). Finally, the residual (RESID) was compared to the one-point impulses employing the residual timing found from the energy waveform's peaks (IT). Also, this excitation of impulses with the residual

timing (1T) was judged in two comparisons to two types of "standard" pulse/noise excitations (1-10 ; 1-14). These two types differ only in the number of LP coefficients used to generate the residual for the pitch estimation routine. Further, these two types of impulse excitations were compared to each other to determine the effect the number of LP coefficients had on the pitch estimation.

The listeners were instructed to choose between the two sentences played for them. If there was almost no discernable difference in quality to the listener, NO DIFFERENCE could be chosen. The results are listed in TABLE 4.1 for four speakers and five listeners.

It should be noted that both the comparisons to the residual excited speech highly favored its quality (85% in #1 and 92.5% in #3). Thus, the major indication was that neither periodic excitation contained enough information to produce as much quality as found from the residual.

The second comparison confirmed the inability of the linear filters to completely separate the two components of the residual. A small precentage preferred the periodic component alone to the combination of the two components (30% to 15%). However, in most cases, there was considered to be no difference between the two. This implies that the addition of the non-periodic component did nothing or, in some cases, corrupted the excitation. Yet, it is assumed that if the periodic component was separated out completely, the addition of the non-periodic component would only improve it.

TABLE 4.1A

Excitation	Preferred A	Preferred B	No Difference
#1) A. Resid B. Per	85.0%	0.0%	15.0%
#2) A. Per B. Per/Non	30.0%	15.0%	55.0%
#3) A. Resid B. IT	92.5%	0.0%	7.5%
#4) A. IT B. 1-10	55.0%	20.0%	25.0%
#5) A. IT B. 1-14	37.5%	30.0%	32.5%
#6) A. 1-14 B. 1-10	57.5%	7.5%	35.0%

TABLE 4.1B

Excitation	Definition
Resid	The Linear Predictive residual signal.
Per	The periodic component resulting from the output of the all-pole IIR linear filter.
Per/Non	The combination of the outputs of the all-pole IIR filter (PER-periodic) and the modified COMB filter (Non-nonperiodic).
IT	The impulse/noise model using residual timing on the pulses.
1-10	The commonly used impulse/noise model where a tenth order inverse filter was used in the pitch estimation routine.
1-14	The commonly used impulse/noise model where a fourteenth order inverse filter was used in the pitch estimation routine.

Table 4.1B Definitions of the excitations being compared in Table 4.1A.

Residual timing on impulses (1T) was compared to random timing for two orders of pitch estimation filters (#4 and #5). A distinct preference (55%) for residual timing was noted over random timing using the 10th order pitch estimation filter (1-10); a slight preference (37%) was shown for residual timing over random timing using the 14th order pitch estimation filter (1-14). The reason for this is further shown in the sixth comparison between the 10th and 14th order pitch estimation filters on impulses with the "random start phase" timing. A major preference was shown for the 14th order pitch estimation filter (57.5%).

These results confirmed the importance of accurate residual timing. Timing based on peaks in the residual corrects some of the error introduced by the "random start phase" of the conventional pulse/noise excitations. Further, there is shown a need to define the shape of the pulses in the periodic component. Since the impulses have no defined pulse shape, information is lost which might be recouped if a model of the shape is determined. The linear filters, though incorrect in their initial use, did indicate a need for models of both components which would add together correctly.

4.2(b) Experiment #2

The second experiment consisted of six comparisons. The impulses employing residual timing (1T) were compared to five-point (5T), ten-point (10T), and twenty-point (20T) pulses

each from the residual. Also, the residual excitation (RESID) was compared to the twenty-point (20T) pulse excitation. This 20-point pulse excitation was compared to the impulse excitation with "random start phase" timing (1-14) and to the periodic component (PER) extracted by the all-pole filter.

The results in TABLE 4.2 indicate that the excitation consisting of 20-point pulses was definitely inferior to the residual (100% favored the residual). However, there was a distinct preference for the 20-point pulses over the impulses with residual timing (31.25%). Since the all-pole filter output came closest in quality to the residual in the first experiment, it was compared to the 20-point pulses. An obvious preference for the periodic component excitation indicates the need for other information than that provided just by the pulses.

In the first two comparisons, the impulses with residual timing (PER) were preferred over the 5-point (5T) and 10-point (10T) pulses (81.25% and 39.6%, respectively). The lack of speech quality from these pulse shapes indicated a need for more spectral information to retain the necessary flat frequency response. The fifth comparison was between the 20-point pulses (20T) and the impulses with the "random start phase" timing using a 14th order LP filter for pitch estimation (1-14). Sixty-two and a half percent preferred the 20-point pulses. That is 20% more than was preferred over the impulses with residual timing. Obviously, the 20-point pulses have residual timing. Thus, this establishes the need for correct timing and a defined pulse shape.

TABLE 4.2A

Excitation	Preferred A	Preferred B	No Difference
#1) A. 1T B. 5T	81.25%	6.25%	12.5%
#2) A. 1T B. 10T	39.6%	27.1%	33.3%
#3) A. 1T B. 20T	22.9%	45.8%	31.25%
#4) A. Resid B. 20T	100.0%	0.0%	0.0%
#5) A. 1-14 B. 20T	20.8%	62.5%	16.7%
#6) A. 20T B. Per	25.0%	72.9%	2.1%

TABLE 4.2B

Excitation	Definition
5T	Five-point pulses from the residual excitation are placed at the peak energy locations.
10T	Ten-point pulses from the residual excitation are placed at the peak energy locations.
20T	Twenty-point pulses from the residual excitation are placed at the peak energy locations.

Table 4.1A Definitions are given for the excitations used in Table 4.2A.

4.2(c) Experiment #3

The third experiment's results are listed in TABLE 4.3. The residual (RESID) was compared to the 20-point pulses from the residual plus the added noise (20T+N). The residual was favored in quality (87.5%) over the pulses plus noise. Yet, the comparison between the 20-point pulse with added noise (20T+N) to the pulses without added noise (20T) shows a distinct preference for added noise (50%). These results indicate the need for both components in the excitation. However, additional information is needed to yield the quality of the residual.

The 20-point pulses taken directly from the residual plus noise (20T+N) were compared to the 20-point pulse per frame plus noise (20T/F+N), as computed in the second method described in Chapter 3. This method chose a sample pulse from the center of the analysis frame to be used as the pulse model. The excitation 20T/F+N repeated this sample pulse at each pulse location and then added the spectrum-shaped noise to it. An overwhelming preference (77.5%) was shown for the individual 20-point pulses. This shows that a sample pulse in each frame does not necessarily define all the pulse shapes in the frame.

Also, the 20-point pulses taken directly from the residual without noise (20T) were compared to the 20-point pulse models with added noise (20TM+N). These pulse models used the third method described in Chapter 3 where a minimum phase FIR filter was used to model the pulse shape used repeatedly in each

TABLE 4.3A

Excitation	Preferred A	Preferred B	No Difference
#1) A. Resid B. 20T+N	87.5%	0.0%	12.5%
#2) A. 20T+N B. 20T	50.0%	15.0%	35.0%
#3) A. 20T+N B. 20T/F+N	77.5%	2.5%	20.0%
#4) A. 20TM+N B. 20T	10.0%	72.5%	17.5%
#5) A. I-14 B. Opt	57.5%	22.5%	20.0%
#6) A. 1T+N B. 20TM+N	52.5%	5.0%	42.5%

TABLE 4.3B

Excitation	Definition
20T+N	The twenty-point pulses taken from the residual at the peak energy locations and used with a mixture of noise.
20T/F+N	One twenty-point sample pulse from each frame is used at the peak energy locations with an added noise component.
20TM+N	A twenty-point model of a pulse is calculated for each frame and is used at the peak energy locations with an added noise component.
Opt	The optimum pulse developed by Sambur, et. al, (1978) is used as the pulse model.

Table 4.3B Definitions are given for the excitations listed in Table 4.3A

frame described by Figure 3.3. Again, a preference for the individual 20-point pulses was shown (72.5%). This indicates that repeated models show more degradation of speech than modeling the pulses does. Thus the models are not inappropriate but that the repetition in each frame is. A better approach would be to model each pulse in the frame with a minimum phase filter and avoid repeating the same shape

There was an added comparison between the impulses with the "random start phase" timing (1-14) and an optimum pulse (OPTI) taken from an experiment discussed in Sambur, et al. (1978). Although his experiment indicated that this pulse shape was superior to the impulse, the results of this comparison imply otherwise. There was a significant preference for the impulses (57.5%) over the optimum pulse. These results reiterated the need for a detailed pulse shape and an overall flat spectrum.

A final comparison was made between the one-point impulses plus noise (1T+N) to the 20-point model plus noise (20TM+N), where both employ residual timing. Here, 52.5% preferred the impulses plus noise. This indicates that these 20-point models show no significant improvement over the generic impulse. Better models would be needed to incorporate the changing pulse shape.

Figures 4.1 and 4.2 show the comparison of speech waveforms. The original speech is shown with the speech generated from four excitations: the residual, the impulses with residual timing, the 20-point pulses, and the 20-point modeled pulses plus noise for a male and a female speaker. All resemble

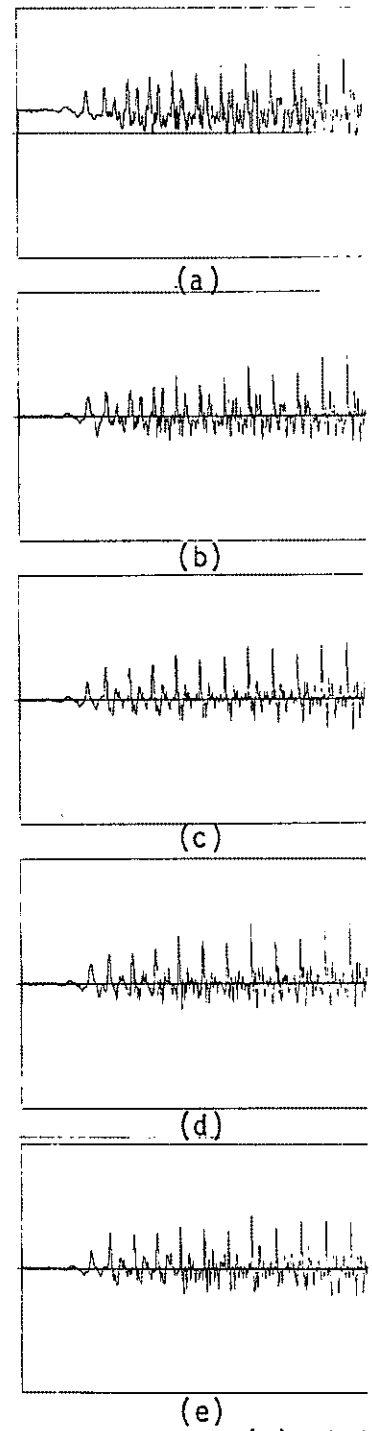


Figure 4.1 Female speaker- (a)Original speech, (b)Residual excited speech, (c)Twenty-point pulse excited speech, (d)Impulses with timing excited speech, (e) Twenty point pulse models excited speech.

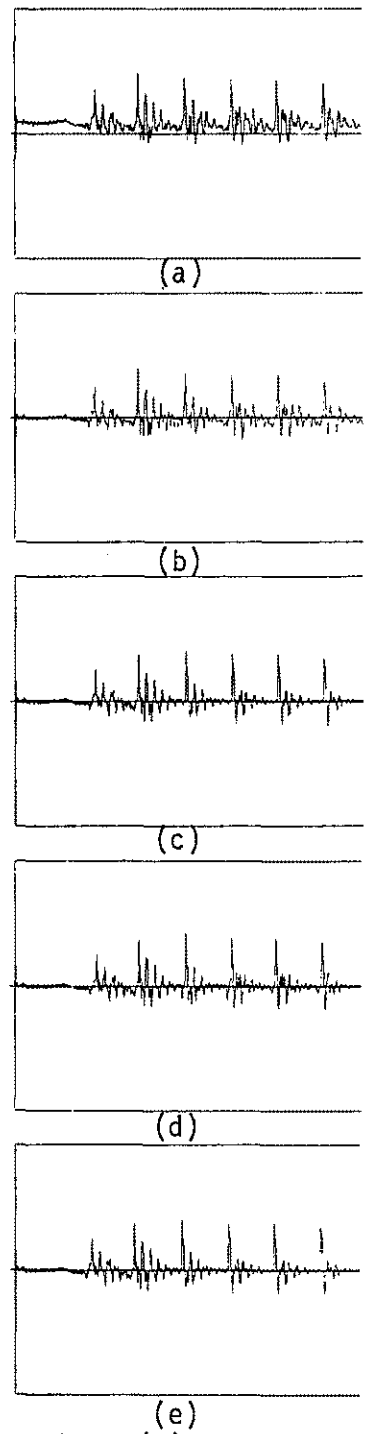


Figure 4.2 Male speaker- (a)Original speech, (b)Residual excited speech, (c)Twenty point pulses excited speech, (d)Impulses with timing excited speech, (e)Twenty point pulse models excited speech.

the original speech closely, yet the residual-excited speech shows the closest resemblance.

4.2 (d) Experiment #4

In the final experiment, the listener was asked to rank in order of preference four sentences. The excitations being ranked were the five and ten-point pulses plus noise (5T+N, 10T+N), the impulses with residual timing plus noise (1T+N) and impulses without noise (1T). TABLE 4.4 lists the final rank between one and four for each excitation, with one as the highest rank and four as the lowest rank in quality.

The highest ranking of 1.8 went to the one-point impulses with residual timing (1T) but without added noise. These results show that the algorithm to determine the added noise is inappropriate for one-point pulses. In preliminary listening tests, the added spectrum-shaped noise improved speech quality when used with 5-, 10-, 20-point pulses. However, this was not the result with the one point pulses. Although the impulses plus noise (1T+N) came in second with a rank of 2.5, it did rank above the 10-point pulses plus noise (2.6) and the 5-point pulses plus noise (3.2).

This experiment reconfirms the results from experiment #2 that a fully defined pulse shape is needed for the periodic component. Otherwise, the missing information degrades speech quality further. A mixture of pulses and noise seems to be a better excitation than pulses alone, except for the case of one point pulses, where an improved algorithm is need to add noise to the one-point pulses.

TABLE 4.4

Excitation	Rank (High 1-Low 4)
1T	1.8
1T+N*	2.5
10T+N*	2.6
5T+N*	3.2

Table 4.4 Ranking of excitations from listening experiment on a scale of one to four where one ranks highest and four ranks lowest. There were four speakers and five listeners.

* The +N denotes that these excitations have the noise mixture with the pulses.

CHAPTER 5

CONCLUSIONS

The following chapter lists the major results and conclusions from this investigation. Applications are discussed and further work is suggested.

5.1 Major Results and Conclusions

Several results were identified in this investigation. The most obvious is the improvement in speech quality due to residual timing on the pulses. A temporal alignment between the impulses in the excitation model and the pulses of the residual excitation was determined to add sufficient information for detectable synthesis improvement. This is shown by the results from experiment #1. First, the quality of speech was increased when using a 14th order filter over a 10th order filter for the pitch estimation to be used on the impulse excitations. Second, using residual timing on the impulses showed an improvement in the speech quality over the impulses which had a random start on the timing. Both of these impulse excitations used the 14th order filter for pitch estimation.

Once the locations of the pulses were found, a pulse shape was also determined. Very high speech quality was derived from the excitations which contained 20-point pulses extracted directly from the residual. However, it was also concluded that pulses made up of too few a number of points lacked enough information to produce improvements in speech quality. A single impulse gave better quality than less "detailed" pulses. Further, it can be concluded that it is difficult to find a pulse shape which works as well as the impulse. A particular example is the comparison in the third experiment between the optimum pulse and the impulse excitations. The optimum pulse was stated by Sambur, et al. (1978), to be superior in synthesis quality over the impulses. The results here prove otherwise. Also, the experiments showed that 5-point, 10-point, and 20-point model pulses were inferior when compared to the impulses. This highlights the difficulty in finding a more effective shape than the impulse. The use of a fixed pulse shape per frame, whether a sample pulse or a pulse model, showed no improvement in quality over the impulses. Perhaps these models should be abandoned in favor of models which take into account the changing pulse shape.

The addition of spectrum-shaped noise improved all cases except the impulses. Here, it was concluded that a better algorithm was needed to determine the spectrum of the noise added to the impulse. The best results came from the 20-point pulses plus added noise, although this model also had the highest data rate of all the models implemented. The addition

of the non-periodic component seemed to only improve the excitation when the pulse was lacking some details.

Finally, a major conclusion was that linear filtering techniques do not separate the periodic and non-periodic components fully. Frequency filtering is dependent upon compatible magnitude and phase spectrums. Thus, separation in the time domain is difficult to achieve. Also, this type of filtering is dependent upon accurate pitch estimates. It is because of these dependencies that these techniques were abandoned.

5.2 Applications and Uses

The primary use of low data rate coding schemes is in digital communications where channel bandwidth is the most limiting factor. Information signals are analyzed for a small number of descriptive parameters. It is these parameters which are transmitted for subsequent reconstruction of the signal. This investigation dealt with ways to improve speech quality, or signal reconstruction, with only a small increase in this number of parameters. For voice communications the quality of the signal can be extremely important. Thus, development of an improved model for the excitation of a Linear Predictive vocoder is one method to ensure quality communications.

Another application is in the field of commercial synthesis. Many products employ the use of speech synthesizers. Repetitious announcements, phone numbers, and

schedules are a few of the places where vocoders have replaced a human voice. The low data rate makes the phrases easily stored and repeatable as often as needed without error. The main drawback is the unnatural quality of the speech. Natural sounding speech is a highly desirable quality to humans who interact with vocoders. An improved excitation model seeks this natural reconstruction of speech.

5.3 Further Work

Further work in this area of investigation would continue along the lines of research stated here. Because of the results identified from this investigation, further work would be to develop a model employing residual timing and individual pulse models for each pulse of the periodic component, plus added noise. Although the 20-point pulses received better results over the 10- and 5-point pulses, it is probable that 15-point pulses could be identified and modeled. The investigation revealed that the peak locations of the pulses were not the middle point of these pulses, and that the pulses were not symmetrical about this location. If the model placed the peak energy at the location point but used an asymmetrical shape, then it is likely that only 15 points would be needed. This model would be minimum phase to best achieve this. An improved algorithm for determining the spectrum of the added noise might be investigated where the final spectrum of the two components was checked to be wide-band. Further, the timing of the pulses would use the precise timing determined by the energy waveform of the LP residual signal.

REFERENCES

B. S. Atal and N. David, "On Synthesizing Natural-Sounding Speech by Linear Prediction," Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, Washington, DC, 1979, pp. 44-47.

B. S. Atal and J. R. Remde, "A New Model of LPC Excitation for Producing Natural-Sounding Speech at low bit rates," Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, Paris, France, 1982, pp. 614-617.

J. Makhoul, R. Viswanathan, R. Schwartz, and A. W. F. Huggins, "A Mixed-Source Model for Speech Compression and Synthesis," J. Acoust. Soc. of Amer., Dec. 1978, Cambridge, Mass., pp. 1577-1581.

J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," IEEE Audio Electroacoust., Vol. AU-20, 1972, pp. 337-367.

C. A. McGonegal, 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. 25, June 1977, pp. 221-229.

L. R. Rabiner and B. Gold, Theory and Application of Digital Signal Processing, Englewood Cliffs, New Jersey, 1975.

L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Englewood Cliffs, New Jersey, 1978.

M. R. Sambur, A. E. Rosenberg, L. R. Rabiner, and C. A. McGonegal, "On Reducing the Buzz in LPC Synthesis," J. Acoust. Soc. of Amer., Vol. 63, March 1978, pp. 918-924.

M. R. Schroeder and B. S. Atal, "Code-excited Linear Prediction (CELP): High-quality Speech at very low bit rates," Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, Tampa, Fla., 1985.

S. A. Zahorian and P. E. Gordy, "Finite Impulse Response (FIR) Filters For Speech Analysis and Synthesis," IEEE Inter. Conf. on Acoust., Speech, and Signal Processing, April 1983, pp. 808-811.

APPENDIX

The following routines were used in the analysis and excitation modeling programs. They are listed in order of discussion in the text.

```
C        OCTOBER    ,1984
C
C        THE FOLLOWING SUBROUTINE IMPLEMENTS A MODIFIED COMB
C        FILTER USING THE DIFFERENCE EQUATIONS-
C
C        1.        Y1(K)=X(K)-X(K-N)
C                                  j*THETA
C        2.        Y2(K)=(Y1(K)-e       * Y1(K-N))
C                                 -j*THETA
C        3.        Y(K)=(Y2(K)-e        * Y2(K-N))
C
       SUBROUTINE COMB(ICHK,IDELAY,RAT,X,Y)
C
C      SUBROUTINE TO IMPLEMENT A MODIFIED COMB FILTER WITH
C      DIMENSIONS STATEMENTS GIVEN:
C
C        X=         INPUT DATA FRAME
C        Y=         OUTPUT DATA FRAME
C        B=         ARRAY FOR STORING PREVIOUS THREE FRAMES
C        IDELAY=    AMOUNT OF DELAY IN COMB FILTER
C        RAT=       RATIO OF DEGREE OF SHIFT TO NUMBER OF DEGREES
C                   BETWEEN ZEROES
C        N=         CONTROL PARAMETER OF FILTER
C        ICHK=      CHECK FOR PREVIOUS PROCESSING
C       NUM=      SELECTS STARTING POINT FOR INPUT ARRAY
C
       COMMON/XXX/ N1,N2,N3,THRESH,IP
       INTEGER N
       DIMENSION B(768),X(1),Y(1)
        REAL RAT,DEG
        N=IDELAY
        IF(ICHK.EQ.0) GO TO 3
        DO 1 I=1,(768-N2)
1       B(I)=0.0
3       CONTINUE
C COMPUTE THE OUTPUT
       NUM=(N1-N2)/2
       DO 4 I=1,N2
4      B(768-N2+I)=X(I+NUM)
       IF(N.EQ.0) GO TO 6
        DEG=RAT*(360.0/FLOAT(N))
       THETA=2*3.1415927*DEG/360
       DO 5 I=1,N2
5      Y(I)=(B(768-N2+I)-(1+2*COS(THETA*N))*B(768-N2+I-N)+
     *  (1+2*COS(THETA*N))*B(768-N2+I-2*N)-B(768-N2+I-3*N))
     *  /4.0 * 0.75
       GO TO 7
6      DO 9 I=1,N2
9      Y(I)=X(I+NUM)
7      CONTINUE
        DO 8 I=1,(768-N2)
8      B(I)=B(I+N2)
        RETURN
       END
```

```
      SUBROUTINE FILT(ICHK,N,A,X,Y)
      DIMENSION X(1),Y(1)
      DIMENSION B2(400)
      REAL A
      COMMON/XXX/ N1,N2,N3,THRESH,IP
      IF(ICHK.EQ.0) GO TO 11
      DO 1 I=1,(400-N2)
1     B2(I)=0.0
11    CONTINUE
      NUM=(N1-N2)/2
      IF (N.EQ.0) GO TO 4
      DO 2 I=1,N2
      Y(I)=(A**N)*B2(I+400-N2-N)+0.5*(X(I+NUM))
2     B2(400-N2+I)=Y(I)
      GO TO 5
4     DO 6 I=1,N2
6     Y(I)=X(I+NUM)
5     CONTINUE
      DO 3 I=1,(400-N2)
3     B2(I)=B2(I+N2)
      RETURN
      END
```

```
      SUBROUTINE FILTER(ICHK,R,VUSD,IPITCH,W)
      DIMENSION R(1),B(415),E(400),H(30),Y(428),W(128)
      COMMON /XXX/ N1,N2,N3,THRESH,IP
      COMMON /XX5/ XMULT,IDEC
C     THE FOLLOWING PROGRAM FIND THE PEAKS OF THE RESIDUAL BY
C     CALCULATING A RUNNING LOW-PASSED 15-POINT ENERGY THROUGH
C     THE ANALYSIS FRAME. DURING THE VOICED FRAMES, FROM THIS
C     SIGNAL THE PEAKS OF THE RESIDUAL ARE DETERMINED AND THREE
C     POINTS OF THIS PEAK ARE SELECTED FOR OUTPUT TO FORM THE
C     EXCITATION. THE ENTIRE RESIDUAL FRAME IS OUTPUTTED DURING
C     UNVOICED FRAMES. A THRESHOLD FOR FINDING THE PEAKS IS
C     DETERMINED BY XMULT. IF IT IS SET TO ZERO THEN THE ENERGY
C     SIGNAL IS OUTPUTTED.
C
      NUM=(N1-N2)/2
      IF(ICHK.EQ.1) THEN
C     RESET LAST TWELVE POINTS
       DO 1 I=1,12
1      B(I)=0.0
C     SET UP FILTER
       H(1)=0.0019474268
       H(2)=-0.00052844197
       H(3)=-0.0052825962
       H(4)=-0.013236498
       H(5)=-0.021686191
       H(6)=-0.025385415
       H(7)=-0.018004123
       H(8)=0.0051981132
       H(9)=0.044532232
       H(10)=0.094369218
       H(11)=0.14394872
       H(12)=0.1805872
       H(13)=0.19409601
       H(14)=H(12)
       H(15)=H(11)
       H(16)=H(10)
       H(17)=H(9)
       H(18)=H(8)
       H(19)=H(7)
       H(20)=H(6)
       H(21)=H(5)
       H(22)=H(4)
       H(23)=H(3)
       H(24)=H(2)
       H(25)=H(1)
      END IF
C     WRITE SQUARES OF INPUT INTO BUFFER
       DO 3 I=1,N1
3      B(I+12)=(R(I))**2
C     CALCULATE RUNNING NINE POINT ENERGY
       DO 6 I=1,(N1-12)
        E(I)=0.0
         DO 5 J=1,25
5        E(I)=E(I)+B(J+I-1)*H(J)
6      CONTINUE
```

```
C    CHECK TO SEND OUT ONLY ENERGY WAVEFORM
        IF (XMULT.EQ.0.0) THEN
         DO 12 I=1,N2
12       W(I)=E(I+NUM)/100.0
        ELSE
C    CHECK FOR VOICED/UNVOICED DECISION
         IF (VUSD.EQ.1.0) THEN
C    INITIALIZE PARAMETERS
          DO 11 I=1,(N1+12)
11        Y(I)=0.0
          E(N1-11)=0.0
          J=1
C         EPEAK2=0.0
C         IPEAK2=0
          INUM=(N1/(IPITCH-10))+1
          DO 200 K=1,INUM
          EPEAK=0.0
          DO 300 I=1,IPITCH
           IF((I+J-1).GT.(N1-11)) GO TO 300
           IF(E(J+I-1).GT.EPEAK) THEN
            EPEAK=E(J+I-1)
            IPEAK=I+J-1
           ENDIF
300       CONTINUE
          Y(IPEAK)=1000.0
          J=J+IPITCH-10
200       CONTINUE
          CALL CLEAN(IPITCH,Y)
          DO 22 I=1,N2
22        W(I)=Y(I+NUM)
         ELSE
          DO 50 I=1,N2
50        W(I)=R(I+NUM)
         END IF
        END IF
C   SAVE LAST TWELVE POINTS OF SQUARES OF INPUT
        DO 7 I=1,12
7       B(I)=B(N1+I)
        RETURN
        END
```

```
      SUBROUTINE CLEAN(IPITCH,Y)
      DIMENSION Y(1)
      COMMON /XXX/ N1,N2,N3,THRESH,IP
      IPEAK2=0
      IPEAK3=0
      DO 1  I=1,N1
      IF (Y(I).EQ.1000.0) THEN
       IPEAK=I
       IF(IPEAK2.EQ.0) GO TO 2
       IF(IPEAK3.EQ.0) GO TO 2
       ISPAC1=IPEAK-IPEAK3
       IF (ISPAC1.LE.(IPITCH+4)) THEN
         Y(IPEAK2)=0.0
         IPEAK2=IPEAK3
       ENDIF
2      IPEAK3=IPEAK2
       IPEAK2=IPEAK
      ENDIF
1     CONTINUE
      RETURN
      END
```

```
      SUBROUTINE NOISE(ICHK,R,VUSD,IPITCH,W)
C*****************************************************************
C     THE FOLLOWING ROUTINE CALCULATES THE ENERGY WAVEFORM FROM
C     THE RESIDUAL AND LOCATES THE PEAKS.  THE LOCATIONS
C     OF THESE PEAKS DETERMINE THE PERIODIC COMPONENT.  THE
C     SPECTRUMS OF THE PULSES AND THE NOISER ARE CALCULATED
C     USING DISCRETE COSINE COEFFICIENTS.  THE PULSE MODEL IS
C     IMPLEMENTED USING A MINIMUM PHASE FIR.  THE NOISE SPEC-
C     TRUM IS SHAPED TO BE COMPLEMENTARY TO THE PULSE USING
C     A LINEAR PHASE FIR.  THESE TWO COMPONENTS ARE ADDED TO
C     MAKE UP THE EXCITATION.
C*****************************************************************
C
      DIMENSION R(1),B(425),E(400),R1(32),X1(32),R2(32),X2(32)
      DIMENSION IPTS(10),A1(5),A2(5),W(1),D(12),H(25),Y(428)
      EQUIVALENCE (E(1),R1(1)),(E(33),X1(1))
      EQUIVALENCE (E(65),R2(1)),(E(97),X2(1))
      COMMON /XXX/ N1,N2,N3,THRESH,IP
      COMMON /XX5/ NX,EE
      NUM=(N1-N2)/2
      IF(ICHK.EQ.1) THEN
C   RESET LAST TWELVE POINTS
       DO 1 I=1,12
1      B(I)=0.0
C   SET UP FILTER
       H(1)=0.0019474268
       H(2)=-0.00052844197
       H(3)=-0.0052825962
       H(4)=-0.013236498
       H(5)=-0.021686191
       H(6)=-0.025385415
       H(7)=-0.018004123
       H(8)=0.0051981132
       H(9)=0.044532232
       H(10)=0.094369218
       H(11)=0.14394872
       H(12)=0.1805872
       H(13)=0.19409601
       H(14)=H(12)
       H(15)=H(11)
       H(16)=H(10)
       H(17)=H(9)
       H(18)=H(8)
       H(19)=H(7)
       H(20)=H(6)
       H(21)=H(5)
       H(22)=H(4)
       H(23)=H(3)
       H(24)=H(2)
       H(25)=H(1)
C
      END IF
C   WRITE SQUARES OF INPUT INTO BUFFER AND CALCULATE ENERGY
       DO 3 I=1,N1
```

```
3       B(I+12)=R(I)**2
C   CALCULATE RUNNING TWENTY-FIVE POINT ENERGY
        DO 6 I=1,(N1-12)
         E(I)=0.0
         DO 5 J=1,25
5         E(I)=E(I)+B(J+I-1)*H(J)
6       CONTINUE
C   CHECK FOR VOICED/UNVOICED DECISION
       IF (VUSD.EQ.1.0) THEN
C   INITIALIZE PARAMETERS
        DO 11 I=1,(N1+12)
11      Y(I)=0.0
        E(N1-11)=0.0
C
C   FIND PEAKS OF RESIDUAL
        J=1
        INUM=(N1/(IPITCH-10))+1
        DO 200 K=1,INUM
        EPEAK=0.0
        DO 300 I=1,IPITCH
         IF((I+J-1).GT.(N1-11)) GO TO 300
         IF(E(J+I-1).GT.EPEAK) THEN
          EPEAK=E(J+I-1)
          IPEAK=I+J-1
         ENDIF
300     CONTINUE
         Y(IPEAK)=1000.0
        J=J+IPITCH-10
200     CONTINUE
C
C   CALL SUBROUTINE TO CLEAN OUT STRAY PEAKS
        CALL CLEAN(IPITCH,Y)
C   SAVE LAST TWELVE POINTS OF SQUARES OF INPUT
        DO 7 I=1,12
7       D(I)=B(N2-12+I)
C
C   STORE LOCATIONS OF EACH PEAK OF MIDDLE 256-PTS.
        NS=(N1-256)/2
        J=1
        DO 23 I=(1+NS),(256+NS)
        IF(Y(I).EQ.1000.0) THEN
         IPTS(J)=I
         J=J+1
        ENDIF
23      CONTINUE
C
C   CALCULATE ENERGY OF 20-POINT PEAKS IN 128-PT FRAME
        NPKS=J-1
        DO 4 I=1,256
4       B(I)=0.0
        DO 21 I=1,NPKS
        DO 28 J=1,20
        L=IPTS(I)-10+J-NS
        IF((L.LE.0).OR.(L.GT.256)) GO TO 28
```

```
        B(L)=R(L+NS)
28      CONTINUE
21      CONTINUE
        EP=0.0
        DO 12 I=65,192
12      EP=EP+B(I)**2
C
C   FIND ENERGY OF FRAME WITHOUT PEAKS
        EN=EE-EP
        IF (EN.LT.0.0) EN=0.0
        EPER=EN/EE
        WRITE(5,201)EPER
201     FORMAT(' PERCENTAGE OF NOISE ENERGY IN THE FRAME IS '
     *  F6.5)
C
C   SELECT A SAMPLE PEAK FROM FRAME
C   CALCULATE ITS SPECTRUM
        ISAMP=NPKS/2
        IF(ISAMP.EQ.0) ISAMP=1
        DO 13 I=1,20
        X1(I)=0.0
13      R1(I)=R(IPTS(ISAMP)-10+I)
        DO 14 I=21,32
        X1(I)=0.0
14      R1(I)=0.0
        CALL FFT842(0,32,R1,X1)
        DO 15 I=1,32
15      R1(I)=(R1(I)**2+X1(I)**2)**0.5
C
C   DETERMINE SPECTRUM OF NOISE
        E4=0.0
        DO 24 I=1,32
24      E4=E4+R1(I)**2
        SUM=0.0
        DO 26 I=1,32
26      SUM=SUM+R1(I)
        ROOT=(SUM/16.0)**2-((E4/8.0)*(1-0.5*EN/EP))
        IF (ROOT.LT.0.0) ROOT=0.0
        XMAX=0.5*(SUM/16.0)+ROOT**0.5
        DO 40 I=1,32
40      IF(XMAX.LT.R1(I)) XMAX=R1(I)
        DO 41 I=1,32
41      R2(I)=XMAX-R1(I)
C
C   SCALE FOR 1/3 POWER
C
        POWER=1.0/3.0
        IF(R1(1).LT.0.000001) R1(1)=0.000001
        IF(R2(1).LT.0.000001) R2(1)=0.000001
        R1(1)=R1(1)**POWER
        R2(1)=R2(1)**POWER
        DO 25 I=2,17
        II=34-I
        IF(R1(I).LT.0.000001) R1(I)=0.000001
```

```
        IF(R2(I).LT.0.000001) R2(I)=0.000001
        R1(I)=R1(I)**POWER
        R1(II)=R1(I)
        X1(I)=0.0
        X1(II)=0.0
        R2(I)=R2(I)**POWER
        R2(II)=R2(I)
        X2(I)=0.0
25      X2(II)=0.0
C
C   CALCULATE DISCRETE COSINE COEFFICIENTS
        CALL DCS(R1,X1,32,A1,5)
        CALL DCS(R2,X2,32,A2,5)
C
C   CALCULATE SMOOTHED ZERO-PHASE SPECTRUM OF PULSE
        CALL SPMAG(32,R1,X1,A1,5)
C   CALCULATE LINEAR PHASE SPECTRUM OF NOISE
        CALL LINPH(A2,R2,5,3,0.0,25)
C
C   DESCALE PULSE SPECTRUM FROM ONE-THIRD POWER
        POWER=3.0
        IF(R1(1).LT.0.000001) R1(1)=0.000001
        R1(1)=R1(1)**POWER
        DO 44 I=2,17
        II=34-I
        IF(R1(I).LT.0.000001) R1(I)=0.000001
        R1(I)=R1(I)**POWER
        R1(II)=R1(I)
        X1(I)=0.0
44      X1(II)=0.0
C
C   CALCULATE MINIMUM PHASE SPECTRUM OF PULSE
        CALL HILB(32,R1,X1)

C
C   PERFORM IDFT ON SPECTRUM OF PULSE
        CALL FFT842(1,32,R1,X1)
C
C   FILTER WHITE NOISE WITH FIR FILTER
        CALL FIR2(ICHK,R2,20,B,E2)
C
C   CALCULATE ENERGY OF 20-POINT PULSES
        EP2=0.0
         DO 27 I=1,NPKS
         DO 35 J=1,20
         K=IPTS(I)-10+J-NS
         IF((K.LE.128-N2/2).OR.(K.GT.128+N2/2)) GO TO 35
         EP2=EP2+R1(J)**2
35       CONTINUE
27       CONTINUE
C
C  SCALE NOISE PROPERLY AND PUT IN FRAME
        NPTS=(256-NPKS*NX)
        E2=E2*(NPTS/256.0)
```

```
        SCALE=((EN/E2)**0.5)*((EP2/EP)**0.5)*0.5
        DO 30 I=1,256
30      B(I)=B(I)*SCALE
C
C FIND ABSOLUTE MAXIMUM POINT OF PULSE
        XPNT=0.0
        DO 42 I=1,20
        IF(ABS(R1(I)).GT.XPNT) THEN
         XPNT=ABS(R1(I))
         IPNT=I
        ENDIF
42      CONTINUE
C
C    WRITE THE PULSES INTO THE BUFFER
         DO 31 J=1,NPKS
         DO 32 I=1,20
         K=IPTS(J)-IPNT+I-NS
         IF((K.LE.0) .OR. (K.GT.256)) GO TO 32
         B(K)=R1(I)
32       CONTINUE
31       CONTINUE
C
        NM=(256-N2)/2
        DO 22 I=1,N2
22      W(I)=B(I+NM)
       ELSE
        DO 50 I=1,N2
50      W(I)=R(I+NUM)
C   SAVE LAST TWELVE POINTS OF SQUARES OF INPUT
        DO 51 I=1,12
51      D(I)=B(N2-12+I)
       END IF
       DO 52 I=1,12
52     B(I)=D(I)
       RETURN
       END
```

```
      SUBROUTINE FIR2(ICHK,H,N,Y,E)
C
C-------------------------------------------------------
C   THIS PROGRAM WILL FILTER WHITE NOISE WHEN PROVIDED
C   WITH THE IMPULSE RESPONSE OF AN FIR FILTER.
C-------------------------------------------------------
C
C     ICHK  =0 FOR FRAMES WHERE NO INITIALIZATION IS
C              NEEDED
C           =1 FOR FRAMES NEEDING INITIALIZATION
C     H     =  ARRAY CONTAINING THE IMPULSE RESPONSE
C              COEEFICIENTS
C     N     =  NUMBER OF FIR IMPULSE COEEFICIENTS
C     Y     =  OUTPUT ARRAY CONTAINING N1 POINTS OF
C              FILTERED WHITE NOISE
C     E     =  ENERGY OF OUTPUT ARRAY
C
      DIMENSION H(1),Y(1),X(300)
      COMMON /XXX/ N1,N2,N3,THRESH,IP
      DO 1 I=1,256
1     Y(I)=0.0
      IF(ICHK.EQ.0) GO TO 20
      DO 2 I=1,N
2     X(I)=0.0
20    DO 3 I=(N+1),(N+256)
      X(I)=RAN(-3,2)
3     X(I)=X(I)-0.5
      DO 4 I=1,256
      DO 5 K=1,N
5     Y(I)=Y(I)+H(K)*X(N+I-K)
4     CONTINUE
      DO 6 I=1,N
6     X(I)=X(I+256)
      E=0.0
      DO 7 I=1,N2
7     E=E+Y(I+(256-N2)/2)**2
      RETURN
      END
```

```
      SUBROUTINE AUTO(N,IP,S,A,XK,INFO)                          90
C     SEPT 5,1977
C     PURPOSE:  CALCULATES AUTOCORRELATION COEFFICIENTS FROM
C               HAMMING-WEIGHTED DATA SEQUENCE.  USES AUTO-
C               CORRELATION COEFFICIENTS TO CALCULATE LP
C               COEFFICIENTS.
C
C               USAGE NOTES
C     ************************************
C     N=        NUMBER OF DATA POINTS
C     IP=       NUMBER OF PREDICTOR COEFFICIENTS
C     S=        DATA ARRAY; LEFT UNCHANGED
C     A=        PREDICTOR COEFFICIENTS
C     XK=       ARRAY OF LP REFLECTION COEFFICIENTS
C     THRESH=   THRESHOLD FOR CALCULATIONS TO BE DONE
C     E=        ENERGY OF DATA
C     INFO=     1 FOR FIRST PASS
C               2 FOR ALL OTHER PASSES
C     XH=       ARRAY OF HAMMING WINDOW WEIGHTS
C     S2=       DATA * HAMMMING WINDOW
C     RR=       AUTOCORRRELATION COEFFICIENTS
C     ************************************
C
C     SUBROUTINES REQUIRED: AUTOLP
C
      COMMON /XXY/ XH
      DIMENSION A(1),XK(1)
      DIMENSION S(1),RR(19),RSC(19),XH(512),S2(512)
    5 FORMAT(10F6.3)
      IPD=IP+1
      GO TO (11,15),INFO
   11 CONTINUE
C     CALCULATE HAMMING WINDOW
      NI=256
      DO 12 I=1,NI
      XH(I)=.54-.46*COS(2.*3.14159265*FLOAT(I-1)/FLOAT(NI-1))
   12 XH(I)=XH(I)*1.4
   15 CONTINUE
      DO 16 I=1,NI
C     USE HAMMING WINDOW
   16 S2(I)=S(I+N/2-NI/2)*XH(I)
      RR(1)=0.0
      DO 10 NN=1,NI
      SUM=S2(NN)*S2(NN)
      RR(1)=RR(1)+SUM
   10 CONTINUE
      DO 20 I=2,IPD
      RR(I)=0.0
      DO 20 NN=I,NI
      NND=NN-I+1
      SUM=S2(NN)*S2(NND)
      RR(I)=RR(I)+SUM
   20 CONTINUE
      CALL AUTOLP(RR,A,XK,IP)
  100 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE AUTOLP(R,A,XK,IP)
C     JAN 13,1978  APRIL 8,1981
C     PURPOSE:  COMPUTES THE LP PREDICTOR COEFFICIENTS AND LP
C               REFLECTION COEFFICIENTS FROM THE AUTOCORRELATION
C               COEFFICIENTS.
C
C               USAGE NOTES
C     ************************************************
C     R=        AUTOCORRELATION COEFFICIENTS
C     A=        LP PREDICTOR COEFFICIENTS
C     IP=       NUMBER OF LP PREDICTOR COEFFICIENTS
C     XK=       LP REFLECTION   COEFFICIENTS
C
C     NOTE THAT IP+1 AUTOCORRELATION COEFFICIENTS ARE USED
C     TO CALCULATE IP LP COEFFICIENTS.
C     ************************************************
C     REFERENCE: MAKHOUL,J., "LINEAR PREDICTION: A TUTORIAL
C     REVIEW,"  PROCEED. OF THE IEEE 63, 561-580, (1975).
C
      DIMENSION R(1),A(1),B(18),XK(1)
   25 FORMAT(2X,F10.4)
      ICOUNT=0
   10 CONTINUE
      E=R(1)
      XK(1)=-R(2)/E
      IF(ABS(XK(1)).LT. .985) GO TO 30
   15 R(1)=1.02*R(1)
      ICOUNT=ICOUNT+1
      IF(ICOUNT .GT. 30) GO TO 200
      GO TO 10
   30 CONTINUE
      A(1)=XK(1)
      B(1)=A(1)
      E=(1.-XK(1)*XK(1))*E
      DO 100 I=2,IP
      I1=I-1
      XK(I)=R(I+1)
      DO 20 J=1,I1
      IJ=I-J
20    XK(I)=XK(I)+B(J)*R(IJ+1)
      XK(I)=-XK(I)/E
      IF(ABS(XK(I)).LT.0.985) GO TO 35
      GO TO 15
35    CONTINUE
      A(I)=XK(I)
      B(I)=A(I)
      DO 40 J=1,I1
      IJ=I-J
   40 A(J)=B(J)+XK(I)*B(IJ)
      DO 50 J=1,I1
   50 B(J)=A(J)
      E=(1.-XK(I)*XK(I))*E
  100 CONTINUE
      GO TO 300
200   CONTINUE
      DO 210 J=I,IP
210   A(J)=0.0
300   CONTINUE
      RETURN
      END
```

```
      SUBROUTINE LPCSYN(N,IP,A,R,S,X,G,IX)
C     AUGUST 18, 1976.
C     AUGUST 17, 1982.
C
C     ************************************************************
C     THIS PROGRAM CAN BE USED TO CALCULATE EITHER THE LPC
C     RESIDUAL SIGNAL(GIVEN THE LPC PREDICTOR COEFFICIENTS
C     AND THE DATA SEQUENCE) OR TO RECONSTRUCT THE ESTIMATE
C     OF THE ORIGINAL SIGNAL(GIVEN THE LPC PREDICTOR COEFFICIENTS,
C     THE SPECTRALLY BROADENED BASEBAND SIGNAL, AND THE GAIN).
C
C     A  =ARRAY OF PREDICTOR COEFFICIENTS.
C     B  =ARRAY OF PREVIOUS IP SIGNAL VALUES.
C     S  =ARRAY OF SIGNAL VALUES FOR CURRENT FRAME.
C     X  =ARRAY OF DUMMY VARIABLES(274 POINTS).
C         EQUIVALENCE (B(1),X(1)),(S(1),X(19)) MUST
C         APPEAR IN THE CALLING ROUTINE.
C     G  =GAIN PARAMETER.
C     IX =1 FOR RECONSTRUCTING THE ORIGINAL SIGNAL.
C        =0 FOR CALCULATING THE RESIDUAL.
C     N  =FRAME LENGTH.
C     IP =NUMBER OF PREDICTOR COEFFICIENTS(LP COEFFICIENTS).
C     ************************************************************
C
      DIMENSION A(1),S(1),X(1),R(1)
      IF (IX.EQ.0) GO TO 24
      DO 20 J=1,N
      S(J)=G*R(J)
      DO 10 K=1,IP
      IPX=18+J-K
10    S(J)=S(J)-A(K)*X(IPX)
20    CONTINUE
      GO TO 45
24    CONTINUE
      DO 40 J=1,N
      R(J)=S(J)
      DO 30 K=1,IP
      IPX=18+J-K
30    R(J)=R(J)+A(K)*X(IPX)
40    CONTINUE
45    CONTINUE
      RETURN
      END
```

```
      SUBROUTINE PEAK7(X,Y,VUSD,PITCH,XPEAK,INFO)
C     ------------------------------------------------------------
C     PROGRAM DESCRIPTION:
C     SUBROUTINE PEAK6 OPERATES ON THE DATA IN ARRAY X AND
C     THEREFORE MAY BE USED FOR PITCH DETECTION IN THE RESIDUAL,
C     THE ACTUAL SPEECH, OR WHATEVER IS DESIRED.
C     THIS VERSION MAY BE MODIFIED TO ANALYZE PITCH SYNCHRONOUSLY
C     BY REMOVING THE COMMENT INDICATORS IN THE NOTED SECTION.
C     ------------------------------------------------------------
C
C
C     ------------------------------------------------------------
C     VARIABLE DESCRIPTION:
C
C     FS = SAMPLING FREQUENCY
C     IDEF    = DEFALUT FOR FRAME LENGTH (EQUAL TO INITIAL
C               FRAME LENGTH)
C     INFO    = 1 FOR FIRST PASS
C             = 2 FOR ALL OTHER PASSES
C     ITRANS = POINTER USED TO INDICATE DATA POINT IN FREQUENCY
C               DOMAIN WHERE LOWPASS FILTER TRANSITION BEGINS
C     L1      = LOWER LIMIT OF THE SEGMENT OF DATA SEARCHED
C               TO FIND PITCH FREQUENCY
C     L2      = UPPER LIMIT OF THE SEGMENT OD DATA SEARCHED
C               TO FIND PITCH FREQUENCY
C     N1      = FRAME LENGTH
C     NSEX    = 1 FOR MALE SPEAKER
C             = 2 FOR FEMALE SPEAKER
C     NUM     = FRAME COUNTER
C     P       = ABSOLUTE PEAK
C     PITCH   = PITCH FREQUENCY OF CURRENT FRAME
C     PITCH2  = PITCH FREQUENCY OF PREVIOUS FRAME
C     VUSD    = 2.0 FOR SILENT FRAME
C             = 1.0 FOR VOICED FRAME
C             = 0.0 FOR UNVOICED FRAME
C     XPEAK   = NORMALIZED PEAKK
C     XH(I)   = ARRAY OF COEFFICIENTS FOR HAMMING WINDOW
C     X,Y     = DATA ARRAY
C        ------------------------------------------------------
C
C
C
         DIMENSION X(1),Y(1)
         DIMENSION XH(512)
         COMMON /XXY/ XH
         COMMON /XXX/ N1,N2,N3,THRESH,IP

         IF (INFO.EQ.2) GO TO 20
C        ------------------------------------------------------
C        INITIALIZATION
C        ------------------------------------------------------
         FS=10000.
         IDEF=N1
         NUM=0
C        FROM N1(FRAME LENGTH) DETERMINE THE SEX OF THE
C        SPEAKER.
         NSEX=2
         IF(N1.GE.400)NSEX=1
```

```
C
C       SET THE LIMITS FOR THE AREA OF SEARCH ACCORDING
C       TO THE SEX OF THE SPEAKER.
        IF (NSEX.EQ.1) GO TO 5
        L1=.0028*FS
        L2=.0101*FS
        GO TO 10
5       L1=.004*FS
        L2=.0201*FS
10      CONTINUE
C
C       CALCULATE HAMMING WINDOW
C
        DO 15 I=1,N1
        XH(I)=.54-.46*COS(2.*3.14159265*FLOAT(I-1)/FLOAT(N1-1))
15      CONTINUE
C
20      NUM=NUM+1
C       ---------------------------------------------------------
C       AUTOCORRELATION:
C       THE VECTOR Z=IX+IY IS REPLACED BY ITS TRANSFORM.
C       ---------------------------------------------------------

C       ZEROS ARE NOW APPENDED FOR SMOOTHING.

        IF (N1.EQ.512) GO TO 30
        N11=N1+1
        DO 25 I = N11,512
           X(I)=0.0
25         Y(I)=0.0

30      CONTINUE
        DO 40 J=1,N1
40         Y(J)=0.0

C       CALCULATE THE DFT.

        IN=0
        N=512
        CALL FFT842(IN,N,X,Y)
C       COMPUTE SPECTRAL MAGNITUDE
        DO 45 K=1,512
           X(K)=X(K)*X(K)+Y(K)*Y(K)
45         Y(K)=0.0

C       LOW PASS FILTERING AT 1000HZ IS NOW DONE.
C
        ITRANS=1000./(FS/512.)
        DO 50 J=ITRANS+5,257
50         X(J)=0.0
        X(ITRANS)=.95*X(ITRANS)
        X(ITRANS+1)=.8*X(ITRANS+1)
        X(ITRANS+2)=.5*X(ITRANS+2)
        X(ITRANS+3)=.2*X(ITRANS+3)
        X(ITRANS+4)=.05*X(ITRANS+4)
C       MIRROR THE SEQUENCE ABOUT POINT 257
        NN=513
        DO 55 I=2,256
```

```
            NN=NN-1
55          X(NN)=X(I)

C       CALCULATE THE IDFT.
        IN=1
        CALL FFT842(IN,N,X,Y)

C   ------------------------------------------------------------
C   PEAK PICKING:
C   THE PITCH PERIOD WILL BE FOUND BY DETERMINING THE ABSOLUTE
C   PEAK IN A TYPICAL PITCH RANGE.  FOR MALES: 50 TO 250 HERTZ,
C   OR 4 TO 20MSEC IN TIME.
C   SIMILARLY, 100 TO 350 HERTZ WILL BE USED FOR FEMALES.
C   PITCH IS DETERMINED USING A SAMPLING PERIOD OF .1MSEC AND
C   THE POINT NUMBER AT THE PEAK.
C   ------------------------------------------------------------
        P=0.0
C
        DO 60 K=L1,L2
           IF (P.GT.X(K)) GO TO 60
           P=X(K)
           PITCH=1./((K-1)/FS)
60      CONTINUE
C
C   ------------------------------------------------------------
C   VOICED/UNVOICED:
C   USING A NORMALIZED AUTOCORRELATION(I.E., IX(1)=1.0, THE
C   SPEECH IS VOICED IF THE ABSOLUTE PEAK IS ABOVE A VALUE
C   OF 0.45.
C   ------------------------------------------------------------
C
        IF (X(1).EQ.0.0) GO TO 70
        XPEAK=P/X(1)
C
C
C       LOWER PEAK THRESHOLD IF PITCH VARIES SMOOTHLY
C
        XLIM=.45
        DEV=(PITCH2-PITCH)/PITCH
        DEV=ABS(DEV)
        IF(DEV .LT. .10) XLIM=.30
C
        IF (XPEAK .LT. XLIM) GO TO 80
C
C       ------------------------------------------------------------
C       OUTPUT:
C       ------------------------------------------------------------
        VUSD=1.0
        WRITE (5,65) NUM,XPEAK,PITCH
65      FORMAT (' FRAME #',I5,'   XPEAK=',F9.5,'   SPEECH IS
     *  VOICED   PITCH=',F9.5,'HERTZ')
        GO TO 90
70      VUSD=2.0
        PITCH=50000.0
        WRITE (5,75) NUM
75      FORMAT (' FRAME #',I5,
     *          '   SPEECH IS UNVOICED(SILENCE)')
        GO TO 90
```

```
80      VUSD=0.0
        PITCH=50000.0
        WRITE (5,85) NUM,XPEAK
85      FORMAT (' FRAME #',I5,
     *          '   XPEAK=',F9.5,'   SPEECH IS UNVOICED')
90      CONTINUE
        PITCH2=PITCH
        RETURN
        END
```

```
      SUBROUTINE LPRES(INFO,D,L,E,Q)
C
C                        USAGE NOTES
C     ------------------------------------------------------
C     SUBROUTINE TO BE CALLED FROM SKELETON PROGRAM
C     FOR LP SYNTHESIS
C
C     D,L,E=   INPUT DATA (RESIDUAL,LP COEFF.,ENERGY)
C     Q=       OUTPUT DATA (SYNTHESIZED SPEECH)
C     N2=      FRAME SIZE
C     IP=      #OF LP COEFFICIENTS (MAX IS 14)
C     LLAST=   ARRAY OF LP COEFFICIENTS FROM PREVIOUS FRAME
C
C     ------------------------------------------------------
C
      COMMON /XX2/ N1,N2,N3,THRESH,IP
      REAL L(1),S(128),LLAST(18),X(146),B(18)
      REAL S2(128),X2(146),B2(18),D(1),Q(128)
      EQUIVALENCE (B(1),X(1)),(S(1),X(19))
      EQUIVALENCE (B2(1),X2(1)),(S2(1),X2(19))
      IF(INFO .EQ. 2) GO TO 58
C     INITIALIZE
      DO 30 I=1,18
      B(I)=0.0
      B2(I)=0.0
   30 LLAST(I)=0.0
      N22=N2/2
      NUM=0
      TYPE 59
59    FORMAT(' MODEL TYPE: EXCITATION USING RESIDUAL'/' 0')
58    NUM=NUM+1
      WRITE(5,*)NUM
56    IF (E.LT.THRESH) GO TO 190
C     COMPUTE CONTRIBUTION TO SIGNAL FROM LAST FRAME
      E1=0.0
      DO 57 I=1,IP
      I2=I+18-IP
57    E1=E1+B(I2)**2
      E1=E1/E
      IF (E1.LT..00001) GO TO 60
      CALL LPCSYN(N2,IP,LLAST,D,S,X,0.0,1)
      GO TO 62
60    CONTINUE
      DO 61 I=1,N2
61    S(I)=0.0
62    CONTINUE
C     COMPUTE ENERGY OF CONTRIBUTION FROM LAST FRAME
      CQX=0.0
      DO 70 I=1,N2
   70 CQX=CQX+S(I)*S(I)
      EW=CQX/E
      IF (CQX .GT. E) GO TO 71
      GO TO 79
   71 EW=CQX/E
      EW=EW**.5
      DO 72 I=1,N2
```

```
   72   S(I)=S(I)/EW
        GO TO 140
C       ABOVE STEPS REDUCE GAIN IN S IF SIGNAL TOO LARGE
   79   CONTINUE
        CALL LPCSYN(N2,IP,L,D,S2,X2,1.0,1)
        CQ2=0.0
        CQ3=0.0
        DO 80 I=1,N2
        CQ2=CQ2+S2(I)*S2(I)
   80           CQ3=CQ3+S(I)*S2(I)
        IF(CQ2 .EQ. 0.0) GO TO 140
        CX=((2.*CQ3)*(2.*CQ3))-(4.*(CQ2)*(CQX-E))
        XG=(-2.*CQ3+CX**.5)/(2.*CQ2)
        DO 90 I=1,N2
   90   S(I)=S(I)+XG*S2(I)
   140  CONTINUE
        DO 160 I=1,IP
        N4=I+N2-IP
        I2=I+18-IP
   160  B(I2)=S(N4)
C       LAST IP VALUES OF SIGNAL ARE SAVED FOR USE BY NEXT FRAME
        DO 170 I=1,IP
   170  LLAST(I)=L(I)
C       LP COEFFICIENTS SAVED FOR USE BY NEXT FRAME
C
        DO 185 I=1,N2
185     IF(S(I) .LT. 0.0) S(I)=S(I)-1.0
C       ABOVE LINE INSURES SMOOTH OVERLOAD
        GO TO 300
   190  CONTINUE
   200  DO 210 I=1,N2
   210  S(I)=0.0
        DO 220 I=1,18
        B(I)=0.0
   220  LLAST(I)=0.0
   300  CONTINUE
        DO 301 I=1,128
301     Q(I)=S(I)
        RETURN
        END
```