
Retrospective Theses and Dissertations

1987

New Excitation Signal for High Quality Linear Predictive Coding Speech Synthesis

William Andrew Pearson
University of Central Florida

 Part of the [Engineering Commons](#)

Find similar works at: <https://stars.library.ucf.edu/rtd>

University of Central Florida Libraries <http://library.ucf.edu>

This Masters Thesis (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of STARS. For more information, please contact STARS@ucf.edu.

STARS Citation

Pearson, William Andrew, "New Excitation Signal for High Quality Linear Predictive Coding Speech Synthesis" (1987). *Retrospective Theses and Dissertations*. 5094.

<https://stars.library.ucf.edu/rtd/5094>

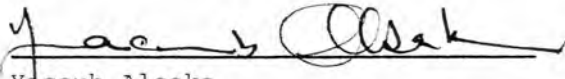
UNIVERSITY OF CENTRAL FLORIDA

OFFICE OF GRADUATE STUDIES

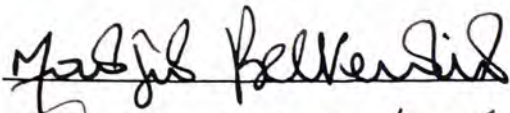
THESIS APPROVAL

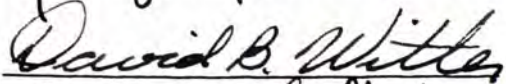
DATE: November 17, 1987

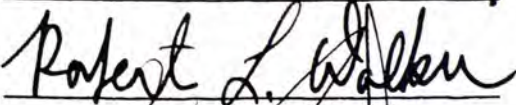
I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION
BY William Andrew Pearson
ENTITLED "A New Excitation Signal for High Quality Linear
Predictive Coding Speech Synthesis"
BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF THE
DEGREE OF Master of Science in Engineering
FROM THE COLLEGE OF Engineering

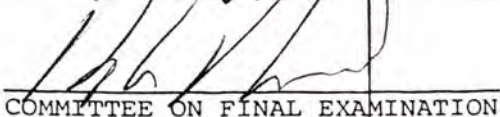

Yacoub Alsaka
Supervisor of Thesis

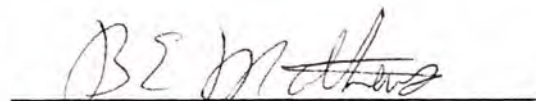
RECOMMENDATION CONCURRED IN:

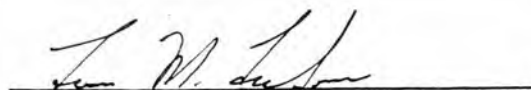

Majid Belkendir


David B. Witter


Robert L. Walker


COMMITTEE ON FINAL EXAMINATION


Bruce E. Mathews
Coordinator of Degree Program


Louis M. Trefonas
Dean of Graduate Studies

A NEW EXCITATION SIGNAL FOR HIGH QUALITY
LINEAR PREDICTIVE CODING SPEECH SYNTHESIS

BY

WILLIAM ANDREW PEARSON
B.S., Auburn University, 1985

THESIS

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

Fall Term
1987

ABSTRACT

The purpose of this thesis is to attempt to improve the quality of synthetic speech by using the best excitation signal during Linear Predictive Coding (LPC) synthesis. This thesis examines the human speech system as a basis for our synthetic speech model. Then it closely examines LPC synthesis, including the mathematical details. One dominant factor in producing natural-sounding and intelligible speech is the excitation signal. For LPC synthesis the excitation signal must have a flat frequency spectrum. A train of impulses separated by the pitch period of the speech has been the standard excitation signal for voiced speech in LPC synthesis. Unfortunately, speech produced using this excitation signal has an unnatural "buzz". For natural-sounding speech, the excitation signal should resemble the glottal volume velocity waveform. The glottal volume velocity waveform is a measure of the excitation that produces natural speech and it does not have a flat frequency spectrum. This raises the question: what

type of excitation signal should be used to produce the most natural-sounding speech possible? To answer this question, we examined six excitation signals that are currently being used in LPC synthesis. We also developed many new excitation signals to be used specifically for synthesizing natural-sounding speech. We experimented with the LPC parameters and these excitation signals to determine the conditions that produced the best speech. Then we compared five of the excitation signals in forced pair trials. We found that our new excitation, LF Impulse excitation, produced speech superior in overall quality (that is naturalness and intelligibility) to the others. We conclude, therefore, that LF Impulse excitation, or an excitation similar to it, should be considered when attempting to produce speech that is both natural-sounding and intelligible with LPC synthesis.

ACKNOWLEDGEMENTS

I would like to thank my committee for their guidance as I worked to complete this thesis. I would also like to thank my company advisor, Mr. David Witter, for his help and friendship. I would like to share my sincere appreciation for the technical advice and the personal time which my committee chairman, Dr. Y. A. Alsaka, has contributed. Finally, I wish to share my love for both my friend, Sherry Rahaim, and my family whose support_ never failed.

TABLE OF CONTENTS

LIST OF FIGURES	vii
CHAPTER	
1. NATURAL SPEECH PRODUCTION	1
Introduction	1
Breathing	2
Producing Sounds	7
Shaping Sounds	12
2. SPEECH SYNTHESIS	20
Types of Synthesizers	21
Practical Requirements	24
Linear Prediction	25
LPC Speech Model	35
Synthesis	44
3. EXCITATION SIGNAL	47
Methods for Constructing the Excitation Signal	48
Excitation Signal Requirements	50
Seven Different Parametric Excitation Signals	52
Conclusion	72
4. TESTING FOR NATURALNESS	74
Hardware	76
Software	76
Other Factors Important to Naturalness	80
Comparison Method	84
Results	86
APPENDIX	
A. LIST OF SOFTWARE	91
B. LISTENER EVALUATION TEST	93

C. USER'S GUIDE TO SPEECH SYNTHESIS PROGRAMS	96
D. HARDWARE	109
E. EXCITATION PROGRAM	114
REFERENCES	125

LIST OF FIGURES

1.	The human vocal apparatus	3
2.	Simple model of the human speech production system	7
3.	Parts of the vocal tract	14
4.	Frequency spectrums of glottal pulses, vocal tract response, and human speech	17
5.	Speech production block diagram	36
6.	Basic speech production model	36
7.	Vocal tract area function	43
8.	Block diagram of a speech synthesizer	45
9.	Three frames of recorded speech	52
10.	FFT of natural speech	53
11.	Single impulse excitation	54
12.	Speech synthesized using single impulse excitation	55
13.	FFT of single impulse excitation	55
14.	Double impulse excitation	57
15.	Speech synthesized using double impulse excitation	57
16.	FFT of double impulse excitation	58

17.	Triple impulse excitation	59
18.	Speech synthesized using triple impulse excitation	59
19.	FFT of triple impulse excitation	60
20.	Fant's glottal flow model	62
21.	First derivative of Fant's excitation	63
22.	Speech synthesized using the first derivative of Fant's excitation	63
23.	FFT of the derivative of Fant's excitation	64
24.	LF excitation	65
25.	Speech synthesized using LF excitation	66
26.	FFT of LF excitation	66
27.	Hilbert transform excitation	68
28.	Speech synthesized using the Hilbert transform excitation	68
29.	FFT of the Hilbert transform excitation	69
30.	LF impulse excitation	71
31.	Speech synthesized using LF impulse excitation	71
32.	FFT of LF impulse excitation	72
33.	Flow of data through software in speech synthesis	77
34.	Flowchart for program Wuyesig	79

35.	Hardware for speech tests	110
36.	Schematic for linear phase lowpass filters	112

CHAPTER 1

NATURAL SPEECH PRODUCTION

Introduction

The purpose of this thesis is to attempt to improve the naturalness of synthetic speech by using the best excitation signal during linear predictive coding (LPC) synthesis. In preparation to meet that goal, we will examine the human speech production system and the model which we will use to simulate that system. Then we will concentrate on a specific part of that model, namely the excitation signal.

The physical processes involved in human speech production are extremely complicated. It would be undesirable, as well as impossible, to synthesize speech by duplicating precisely the processes that result in natural speech. What we want to do instead is to find a simplified model for the actual system. This model must be simple enough to be used practically, but it must be complicated enough to produce good quality speech. Finding a model for a system is a common engineering problem. The first

step is to examine the system and attempt to determine the most important factors affecting its output. Thus our first step will be to examine the important physical processes involved in producing natural speech. Once an understanding of the system is gained, a model to simulate it can be proposed which incorporates the most important factors of the system. Our second step will be to examine the models that have been proposed to synthesize speech. We will briefly examine the different models with emphasis on linear predictive coding. Let us examine natural speech production.

Breathing

The major parts of a person's vocal apparatus are shown in Figure 1 (Oppenheim 1978). The vocal tract is a nonuniform acoustic tube about 17 cm. in length (for a male). At one end of the vocal tract is the glottis (the opening between the vocal cords) and at the other end are the lips. The shape of the tract is determined by the placement of the lips, jaw, tongue, and velum. Another cavity, the nasal tract, can be coupled with the vocal tract by the trapdoor action of the velum. The nasal tract is about 12 cm.

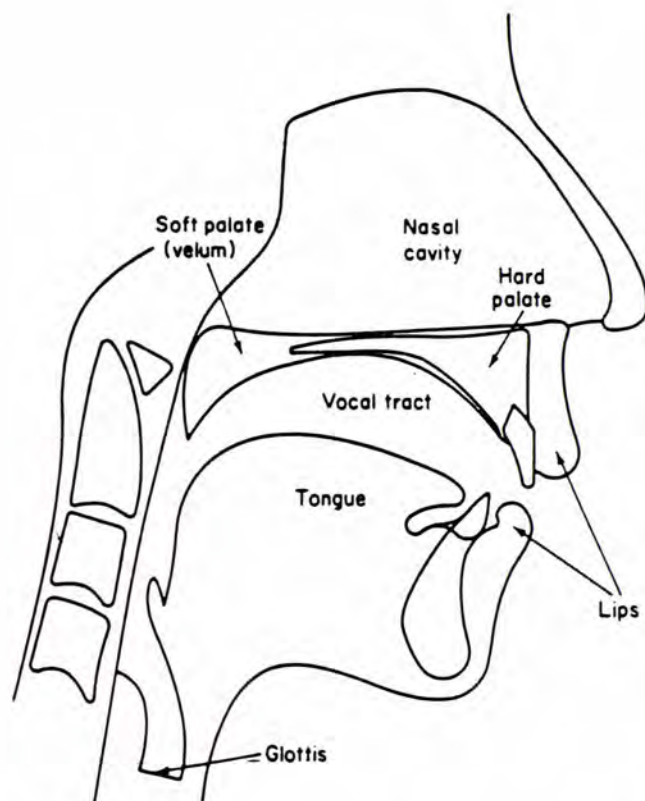


Figure 1. The human vocal apparatus (modified from Borden and Harris 1980).

long in a man. It begins at the velum and ends at the nostrils (Flanagan 1970). Speech is begun by forcing air from the lungs up through the glottis. This pressure is referred to as subglottal pressure. The air travels through the glottis and vocal tract and is radiated into the air. The sounds that a person perceives are variations in pressure which have

radiated out from the source of the sound. This chapter will examine four main physiological functions necessary for speech:

1. Respiration - breathing (there must be a stream of air for speech to be produced).
2. Phonation - converting the air supply into sound.
3. Articulation - movements of the tongue, pharynx, palate, lips, and jaw to "shape" speech sounds.
4. Resonance - acoustic response of the air within the vocal tract to a source of sound.

Shaping Airflow

In English speech, sounds are a consequence of modifying airflow from the lungs. The speaker produces a stream of air and then shapes it to produce sounds for a listener. Limits are set by the speech mechanisms. A person has only a few movable parts with which to originate sounds. The parts are the vocal folds, tongue, jaw, lips, and soft palate. The only three cavities to use as resonators are the oral, pharynx, and nasal cavities. Yet speakers create a multitude of sounds (which are known as phonemes) (Borden and Harris 1980).

Inhalation

To inhale, a person contracts the diaphragm. The diaphragm is beneath the lungs and as it is lowered the volume inside the lungs increases. At the same time the intercostal muscles, which run between the ribs, contract to elevate the ribs. Because of the increased volume within the lungs, the pressure is lower than the atmospheric pressure outside. In order to keep the pressure equal, air from the outside moves to the area of lower pressure within the lungs.

Exhalation

At a high lung volume, a large muscle force is required to maintain the volume. If the muscles are relaxed, the air will rush out due to three passive forces: the elastic recoil of the lungs and rib cage, torque from the untwisting of the cartilages next to the sternum, and gravity which tends to lower the rib cage. These three forces are sufficient to exhale air from the lungs.

Sustained Voicing. The passive forces of exhalation are not adequate by themselves to sing a note or to speak. Exhalation during singing then differs from that during quiet breathing, and exhalation during

speech differs from both. In order to maintain a constant pressure to produce a note sung at a constant intensity, the passive recoil forces are used as a background to which is added active muscle contractions. If a singer permitted expiratory forces to act unaided, the lungs would collapse suddenly and the note could not be sustained. The active muscles slow down the outflow.

Running Speech. The inspiratory muscles (muscles used for inhalation) check the rate of exhalation during running speech as they do in sustaining a tone. During running speech, intensity is constantly changing because certain sentences, phrases, words, and syllables are given emphasis. In order to increase the intensity of the speech sound, the speaker must increase subglottal pressure. The abdominal muscles are used for added expiratory force in heavily stressed utterances or for long utterances. Stressed syllables are produced by possible increases in three factors: duration, frequency, and intensity. Intensity is controlled by subglottal pressure. But a subglottal air pressure increase is not usually associated with particular phonemes in English, so

during running speech the respiratory system continues to supply a fairly constant pressure for a given utterance. It is the opening and closing at the glottis and in the vocal tract which alters the airflow and air pressure as we measure them at the mouth for different speech sounds.

Producing Sounds

Air exhaled from the lungs is the power supply for speech, but it is the upper airways that change this air supply into sounds for speech. Speakers use two methods of converting the air into sounds for speech. The first method involves using the air pressure to set the elastic vocal folds which lie in

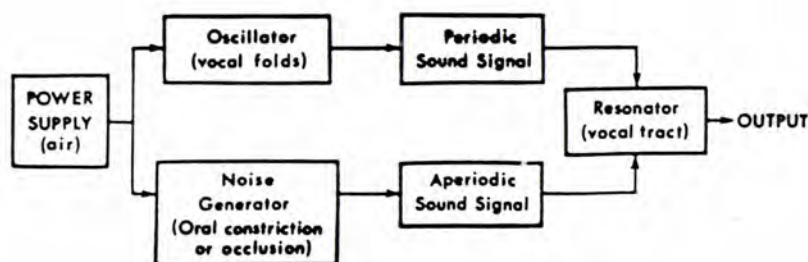


Figure 2. Simple model of the human speech production system (Borden and Harris 1980).

the larynx into vibration, producing a periodic sound. The second method does not involve vibration of the vocal folds. It involves allowing air to pass through the larynx into the vocal tract, where various modifications of the airstream result in noise: bursts, hisses, or combinations of these aperiodic sounds (see Figure 2). The results of the first method are voiced sounds and the second method produces unvoiced sounds.

Theory of Phonation

A person's vocal folds are elastic bulges of tendon, muscles, and mucous membrane. They can be made thick or thin and short or long. They can be opened or closed together. Their tension and elasticity can be varied. In running speech, all these adjustments occur rapidly. When the vocal folds are together and vibrating, they are in the voicing mode. In the myoelastic aerodynamic theory, the key word is aerodynamic. The vocal folds are activated by the airstream from the lungs rather than vibrating by themselves. Myoelastic refers to the way in which the muscles change their elasticity and tension to effect changes in the frequency of vibration.

The number of times the vocal folds open and close per second is the frequency of vocal fold vibration. This frequency directly determines the lowest frequency (fundamental frequency) of the sound which is produced. Mens' voices have an average fundamental frequency of approximately 125 Hz. Womens' voices have an average pitch closer to 200 Hz. Size of the vocal folds is one determinant of the fundamental. The larger the vibrating mass, the lower the frequency. Men have larger vocal folds than women. Given vocal folds of a particular size, however, a person can increase the frequency of vibration appreciably by lengthening and tensing the folds, thus decreasing the effective mass (Borden and Harris 1980).

Vocal Fold Movements

During running speech, the vocal folds are separated for voiceless speech sounds, such as the consonants /s/ and /t/. They are brought together for voiced sounds, such as the vowels /u/ and /i/, and they are less firmly brought together for voiced consonants, such as /z/ and /d/. The vocal folds at rest are apart, creating a v-shaped glottal space.

For voiceless consonants the vocal folds merely open or abduct in order to allow sufficient airflow. In order to produce the voiced sounds of speech, the folds must be adducted (brought together) or nearly so because open folds cannot be set into vibration.

Subglottal Air Pressure

When a large enough subglottal air pressure is applied to the lower part of the vocal folds, they are forced open. During each opening, a tiny puff of air escapes and this set of pulses sets up a pressure wave at the glottis which is audible. As the top part of the vocal folds opens, the bottom part closes. Thus there is a wave-like motion of the folds during vibration.

Bernoulli Effect

The negative pressure which occurs as air passes through the folds pulls them back together. This action is called the Bernoulli effect. The Bernoulli effect is based on the observation that when a gas or liquid runs through a constricted passage, the velocity increases. The Bernoulli principle states

that this increase in velocity results in a drop in pressure perpendicular to the direction of flow (Borden and Harris 1980).

Vocal Fold Vibration

Each cycle of vocal fold vibration during voicing is caused by both the subglottal air pressure and the Bernoulli principle. The subglottal pressure opens the folds and then the Bernoulli principle accounts for a sudden drop in pressure against the inner sides of each fold and pulls them towards each other. The folds elasticity not only permits them to be blown open, but the elastic recoil force works along with the Bernoulli principle to close the folds for each cycle of vibration.

The vocal folds move in a fairly periodic way during voicing. Like all complex periodic sounds, the sound wave contains harmonics. The fundamental frequency is the number of glottal openings per second. We never hear the sound of vocal fold vibration, because by the time it reaches the lips of the speaker, it has been changed by the vocal tract. It is a characteristic of the human voice that the

higher harmonics have less intensity than the lower ones, so that the emphasis is on the lower frequencies.

Summary

Producing speech is a dynamic process. It varies during running speech in intensity, frequency, and quality. Speech is a rapidly varying stream consisting of silence, periodic sounds, and noises. Producing speech must be coordinated with breathing. To take a breath for speech, the glottis opens quickly, and when the vocal folds adduct for voicing, the action is simultaneous with exhalation.

Shaping Sounds

The Vocal Tract

Included in the vocal tract are all of the air passages above the larynx from the glottis to the lips. There are three main cavities which resonate: the pharyngeal cavity, the oral cavity, and when it is open, the nasal cavity (Borden and Harris 1980).

Sounds Produced

The results of filtering the periodic wave produced at the glottis through the vocal tract are

the speech sounds which we know as vowels, semivowels, and nasals. The cavity variations produce resonance changes which make the sounds unique. The speech wave is periodic because of the vibrations of the vocal folds.

Transient sounds created by impeding the airstream and then releasing built-up pressure are called plosives. Forcing the airstream through a constriction results in noisy turbulence. This type of sound is generally called a fricative. These constrictions can be at the glottis, throat, or the oral cavity. Speech sounds can be combined in a variety of ways. The most resonant, open tract sounds, are the vowels, diphthongs, and semivowels. The less resonant, constricted vocal tract sounds are the nasals, stops, and fricatives.

Parts of the Vocal Tract

Pharynx. The pharynx is a tube of muscles which makes up the bottom of the vocal tract. Contraction of the constrictor muscles around the pharynx narrows the pharyngeal cavity. The nasal, oral, and laryngeal cavities open into the pharyngeal cavity (see Figure 3) (Borden and Harris 1980).

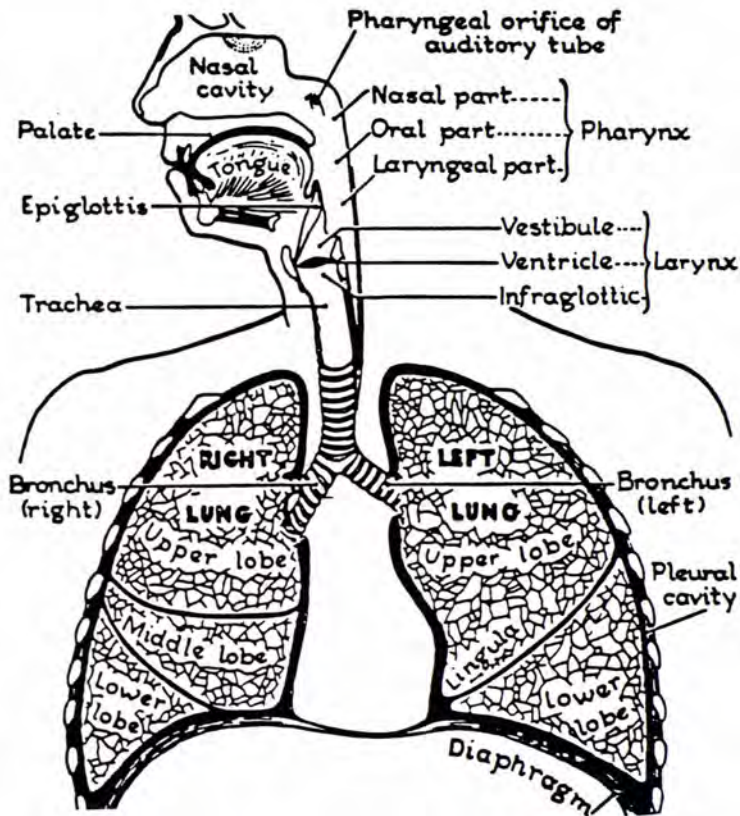


Figure 3. Parts of the vocal tract (modified from Borden and Harris 1980).

Oral Cavity. The oral cavity is the space between the upper and lower teeth. The incisors are the most important teeth for speech. They are the flat-edged

teeth in the front of the mouth. Incisors are used with the lips and tongue to create a constriction for such sounds as /f/ and /s/. The roof of the oral cavity consists of the hard palate and the soft palate or velum.

The Velum. Most of the soft palate consists of a broad muscle. This muscle's function is to elevate the soft palate, thus closing the entrance to the nasal cavities above.

The Tongue. The bottom of the oral cavity is the muscle known as the tongue. The tongue can be moved in different directions for different speech sounds.

The Lips. There are many facial muscles, most of which are intertwined with the major lip muscle, the orbicularis oris, which circles the lips. The lips are closed for sounds such as /u/ or /w/.

Theory of Vowel Production

Tube Resonance. The vocal tract approximates a tube closed at one end and open at the other when vowels are spoken. The vocal folds are essentially closed during voicing and the lips are open. It is commonly

known that the lowest frequency of resonance will have a wavelength (w) 4 times the length of the tube. A male vocal tract has a length of about 17 cm. So the wavelength for the lowest resonant frequency is $4 \times 17 = 68$ cm. The velocity of sound in air (c) is 344 m/s. Since $f = c / w = 34,400 / 68 =$ approximately 500, the lowest resonant frequency is about 500 Hz. The tube also resonates at odd multiples of this frequency.

Male Vocal Tract Resonance. The sound produced by the vibrations of the vocal folds is changed by the resonant response of the vocal tract. The changes can best be understood by comparing the sound at its source (the glottis) with the final output at the lips. The changes that occur are due to the effects of transmission through the vocal tract. The exact nature of the changes can best be appreciated by contrasting the Fourier spectra. The spectrum of the sound source consists of a fundamental frequency and harmonics (see Figure 4). The harmonics diminish in intensity as they increase in frequency. The middle spectrum in Figure 4 shows the resonant frequencies of a neutral vocal tract (500, 1500, 2500 Hz.). These

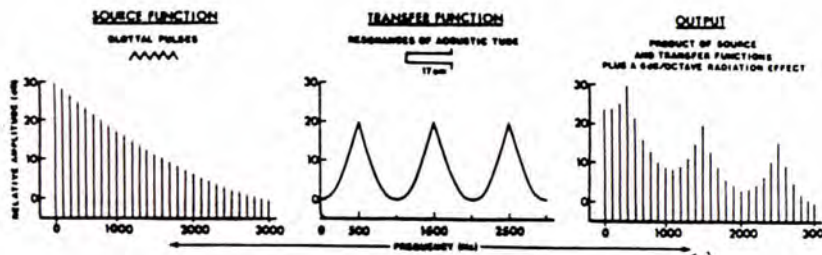


Figure 4. Frequency spectrums of glottal pulses, vocal tract response, and human speech (modified from Borden and Harris 1980).

are the frequencies at which the air in the tract will vibrate maximally in response to a complex sound. The sound which emerges at the end of the tract has the same harmonics as the source, but the amplitude of the harmonics has been modified, altering the quality of the sound.

Vowels. Each vowel is a combination of different resonances than the other vowels. When the vocal tract is shaped like a tube of uniform cross-section, its resonances are odd multiples of the lowest resonance. When the shape is changed and parts of the tract are constricted, the resonances change frequency and they lose their simple relationship to each other.

Tense-Lax Vowels. Certain vowels naturally last longer than others and are made by the tongue stretching farther. These vowels are termed 'tense vowels. Shorter vowels can occur only in closed syllables (syllables ending in consonants). These are named lax vowels since they are made with less tongue movement.

Diphthongs

Diphthongs are vowels with a changing resonance. Those tense vowels which when held result in a changing vocal tract are diphthongs.

Semivowels

The sounds /w/, /j/, and /v/ are termed semivowels because, like vowels, they require considerable resonance. The vocal tract is relatively open as for vowels and diphthongs, but the semivowels are considered consonants.

Nasals

When we considered vowels and diphthongs, we looked merely at the source of the sound and the resonances of the vocal tract. For consonants the constrictions also produce decreased energy in certain frequency ranges (antiresonances). Nasal resonance is

mandatory for the production of /m/ and /n/, so the velum is low, leaving the entrance to the nasal cavity open.

Sounds Produced in the Vocal Tract

Plosives. The oral cavity is closed at some point for plosives such as /p/ and /t/. The closure creates a rapid growth of air pressure within the oral cavity, which is suddenly released by relaxing the blockage. The audible burst of air which results is impossible to prolong. It is transient. Stops can be either voiced or voiceless.

Fricatives. Many sounds can be originated in the vocal tract by sending the air flow (either voiced or voiceless) through constrictions formed in the tract. The pressure must be high enough and the constriction must be small enough to produce friction. The fricative sounds of speech depend upon compressing a continuous airflow through a narrow passage (like the whine of steam from a kettle). Unlike plosives, fricatives can be prolonged.

CHAPTER 2

SPEECH SYNTHESIS

The purpose of this chapter is to survey the different types of speech synthesizers and then give a detailed description of linear predictive coding (LPC) synthesis, since it is the type that we will use to try to synthesize natural sounding speech. We first discuss some of the advantages and disadvantages of each type of synthesizer. Next will be a discussion of the practical requirements of a working speech synthesis system. Next, since our particular area of interest is LPC synthesis, is a description of the general LPC model including the analysis method used to compute the predictor coefficients. From the general LPC model we develop a particular model for the speech system. This includes a determination of the number of predictor coefficients that are required and the inclusion of several other parameters in addition to the predictor

coefficients. Finally, there is a brief discussion of how a speech signal is synthesized using the parameters computed during analysis.

Types of Synthesizers

For many applications, it is desirable to be able to convert English from a form that can be efficiently stored on a computer to natural and intelligible sounding speech. This conversion, called speech synthesis, is accomplished in several different ways depending on the form in which the information is stored. In order to synthesize speech, enough information about the speech signal must be stored to recreate the signal or a close approximation to it.

The many synthesizers now in use can be broken down into four general types. Listed on the next page the types are characterized by the information about the speech signal which is stored. Each type of synthesizer has its own advantages and disadvantages.

Recorded Speech

This is by far the simplest and easiest way to synthesize speech. The signal may be simply recorded and played back. The speech waveform may also be digitized before being stored. Simply storing the

speech waveform works very well when only a small number of words or phrases must be synthesized. However, it requires much more memory than other methods and sounds unnatural when separate words or letters are strung together. Four types of synthesizers and the information they store to describe the speech signal are shown below:

1. recorded speech - the sampled signal is stored directly.
2. articulatory - parameters describing the mechanical motions and resulting volume and pressure changes in the lungs and vocal tract are stored.
3. formant - information about the frequency spectrum of the speech signal (specifically, the location of resonant frequencies) is stored.
4. LPC - parameters describing the speech waveform in terms of the transfer function of the vocal tract and characteristics of the excitation source are stored.

Articulatory

This method is one of the most complicated being used to synthesize speech. But since it attempts to

model the physical processes that produce human speech, it should have the potential to produce the most natural-sounding speech of all the synthesizers and provide the most insight into how speech is actually produced. Because it is complicated, this method has not been the most successful at synthesizing understandable speech and has been used mostly for academic reasons rather than practical systems (Klatt 1980).

Formant

Until recently this was the dominant method for synthesizing speech because it stored relatively simple information about the frequency spectrum of the signal, but was able to produce natural and understandable speech. Unfortunately, applying spectral analysis to speech signals has some limitations resulting from the peculiar properties of the speech wave, such as non-stationarity (Atal and Hanauer 1971).

Linear Predictive Coding

LPC is an increasingly popular method for synthesizing speech which represents the signal in terms of parameters describing the time-varying vocal

tract filter transfer function and characteristics of the excitation source function. Since this method models the time domain signal, it can more accurately represent rapidly changing events than a frequency domain model such as formant synthesis.

Practical Requirements

Practical speech synthesis systems can be characterized by two main requirements. First, the number of words which are required to be synthesized and second, the degree to which the synthesized speech must sound like natural human speech.

Number of Words

This requirement can range from unlimited vocabulary to a small working vocabulary. Unlimited vocabulary speech synthesis could be used to produce a talking encyclopedia, while an example of a small vocabulary system is the telephone information system which includes the words {the, number, is, one, two, ...,nine} and combines them to give a seven digit telephone number (Kaplan and Lerner 1985).

Degree of Naturalness

This requirement can range from speech that is indistinguishable from that of a human to speech that makes no attempt to flow together but is still understandable. In the first example described above, the talking encyclopedia would require very natural sounding speech so that the meaning of a passage would be retained and so that hearing long passages would not irritate the listener. The telephone information system, however, may function very well even though it produces speech that sounds obviously unnatural. The purpose of this paper is to explore how the most natural sounding speech can be obtained in LPC synthesis.

Linear Prediction

The analysis of the behavior of dynamic systems is of concern in many fields. Most of the analysis of the outputs of dynamic systems is done under the title "time series analysis." The majority of the work on time series analysis has been done by statisticians. One aspect of time series analysis is linear prediction.

A continuous signal, $s(t)$, is sampled to give a discrete signal $s(nT)$, when using time series analysis. The variable t is time, n is an integer variable, and T is the constant sampling interval. The sampling frequency $f = 1/T$. For convenience $s(nT)$ is usually abbreviated as $s(n)$. Time series analysis can be used for system modeling. By developing a parametric model for the behavior of a signal-prediction, control, and data compression can be achieved (Makhoul 1975).

An extremely useful model being used extensively is that where a signal $s(n)$ is the output of some system with some unknown input $u(n)$. The system obeys the relationship below.

$$s(n) = - \sum_{k=1}^p a(k)s(n-k) + G \sum_{l=0}^q b(l)u(n-l), \quad b(0)=1 \quad (1)$$

The parameters for this system are $a(k)$, $b(l)$, and the gain G . The output $s(n)$ is a linear sum of past outputs and present and past inputs. So the signal can be predicted from past outputs and inputs. This is why the model is called linear prediction.

All-Pole Model

We can rewrite equation (1) in the frequency domain by taking its z transform. Then, by rearranging, the transfer function of the system $H(z)$ can be obtained.

$$H(z) = \frac{S(z)}{U(z)} = G \frac{1 + \sum_{l=1}^q b(l)z^{-l}}{1 + \sum_{k=1}^p a(k)z^{-k}} \quad (2)$$

$S(z)$ in (3) is the frequency domain representation of $s(n)$. $U(z)$ is the frequency domain representation of $u(n)$.

$$S(z) = \sum_{n=-\infty}^{\infty} s(n)z^{-n} \quad (3)$$

Two particular cases of this model are the all-zero model, where all the $a(k)$'s are equal to zero and the $b(l)$'s are nonzero, and the all-pole model, where all the $b(l)$'s are equal to zero and the $a(k)$'s are nonzero. The all-zero model is called the moving average (MA) model by statisticians, and the all-pole model is the autoregressive (AR) model. The pole-zero

model is therefore the autoregressive moving average (ARMA) model. This paper will use only the all-pole model since it is used almost exclusively in linear prediction of speech. This is due to the all-pole model's ability to accurately model speech and to the model's simplicity.

$S(n)$ for the all-pole model is a linear sum of previous values and the input $u(n)$ (G is a gain factor).

$$s(n) = - \sum_{k=1}^p a(k)s(n-k) + Gu(n) \quad (4)$$

This causes $H(z)$ in (2) to reduce to an all-pole transfer function.

$$H(z) = \frac{G}{1 + \sum_{k=1}^p a(k)z^{-k}} \quad (5)$$

For any given signal $s(n)$, it is necessary to find the predictor coefficients $a(k)$'s and the gain G in some way. The derivation of equations to determine these parameters can be done in several ways. A least squares minimization will be used here (Makhoul 1975).

We will assume first, for clarity, that $s(n)$ is a deterministic signal and then that $s(n)$ is a sample from a random process.

Analysis Method

In many cases, such as speech analysis, the input $u(n)$ is completely unknown. Therefore, the signal $s(n)$ can only be approximated from a linear weighted summation of previous values. We call this approximation $\tilde{s}(n)$ and define it as follows:

$$\tilde{s}(n) = - \sum_{k=1}^p a(k)s(n-k) \quad (6)$$

The error factor included in $\tilde{s}(n)$ is

$$e(n) = s(n) - \tilde{s}(n) = s(n) + \sum_{k=1}^p a(k)s(n-k) \quad (7)$$

The error $e(n)$ is sometimes called the residual. For our method, the least squares approach, the coefficients $a(k)$'s are found by minimizing the squared error with respect to each coefficient. First, assume that $s(n)$ is a known signal (Makhoul 1975).

Deterministic Signal. Let the total squared error be as shown below.

$$E = \sum_n [e(n)]^2 = \sum_n [s(n) + \sum_{k=1}^p a(k)s(n-k)]^2 \quad (8)$$

The limits of the summation in (8) are important, but for simplicity let us minimize E before defining these limits. The minimization is accomplished by making

$$\frac{\partial E}{\partial a(i)} = 0, \quad 1 \leq i \leq p \quad (9)$$

The following set of equations can be derived from (8) and (9).

$$\sum_{k=1}^p a(k) \sum_n s(n-k)s(n-i) = - \sum_n s(n)s(n-i), \quad 1 \leq i \leq p \quad (10)$$

Equations (10) are called the normal equations. It can be seen that (10) forms a set of p equations with p unknowns which can be solved for the a(k) parameters to minimize E in (8) (Makhoul 1975).

To find this minimized error, we expand (8) and substitute in (10) as described below.

$$E_{\min} = \sum_n [s(n)]^2 + \sum_{k=1}^p a(k) \sum_n s(n)s(n-k) \quad (11)$$

The range of summation over n in (8), (10), and (11) can be specified in two ways. This will develop into two different means for estimating the coefficients.

The first method is the autocorrelation method. In this method the error is minimized over all n from $-\infty$ to ∞ . Equations (10) and (11) are rewritten as shown below.

$$\sum_{k=1}^p a(k)R(i-k) = -R(i), \quad 1 \leq i \leq p \quad (12)$$

$$E_{\min} = R(0) + \sum_{k=1}^p a(k)R(k) \quad (13)$$

Where $R(i)$ is the autocorrelation function defined below.

$$R(i) = \sum_{n=-\infty}^{\infty} s(n)s(n+i) \quad (14)$$

However, $s(n)$ is actually known for just a finite time in the real world. A way to overcome this is to multiply the signal by a window $w(n)$ to create a signal $s'(n)$ that is zero except within a range $0 \leq n \leq N-1$ as follows:

$$s'(n) = \begin{cases} s(n)w(n), & 0 \leq n \leq N-1 \\ 0, & \text{otherwise} \end{cases} \quad (15)$$

This changes the autocorrelation function, $R(i)$, to

$$R(i) = \sum_{n=0}^{N-1-i} s'(n)s'(n+i), \quad i \geq 0 \quad (16)$$

The particular window function chosen is important and there are many different windows which can be used. The choice depends on which characteristics of the signal are most important.

The second method is called the covariance method. Here the error E in (8) is minimized over a finite time such as $0 \leq n \leq N-1$. This causes equations (10) and (11) to be rewritten as follows:

$$\sum_{k=1}^p a(k)\Phi(k,i) = -\Phi(0,i), \quad 1 \leq i \leq p \quad (17)$$

$$E_{\min} = \Phi(0,0) + \sum_{k=1}^p a(k) \Phi(0,k) \quad (18)$$

The covariance function is defined as follows:

$$\Phi(i,k) = \sum_{n=0}^{N-1} s(n-i)s(n-k) \quad (19)$$

Random Signal. Given that $s(n)$ is a sample of a random process, it follows that $e(n)$ in (7) is a sample of a random process. For a random signal we work with its expected value. So if the expected value is denoted by EV , then the expected value of the squared error is

$$E = EV[e^2(n)] = EV[s(n) + \sum_{k=1}^p a(k)s(n-k)]^2 \quad (20)$$

When we use (9) in the above equation the result is

$$\sum_{k=1}^p a(k)EV[s(n-k)s(n-i)] = -EV[s(n)s(n-i)], \quad 1 \leq i \leq p \quad (21)$$

Equation (24) shows the minimum average error.

$$E_{\min} = EV[s^2(n)] + \sum_{k=1}^p a(k)EV[s(n)s(n-k)] \quad (22)$$

To evaluate the expected value we must examine two cases (Makhoul 1975).

The first case is a stationary process. The expected value of $s(n)$ is

$$EV[s(n-k)s(n-i)] = R(i-k) \quad (23)$$

It can be seen that (21) and (22) now reduce to (12) and (13).

The second case is a nonstationary process. The expected value of $s(n)$ is

$$EV[s(n-k)s(n-i)] = R(n-k, n-i) \quad (24)$$

$R(n, n')$ is the nonstationary autocorrelation between times nT and $n'T$. For simplicity we will take the case at time $nT = 0$. This case produces the equations below.

$$\sum_{k=1}^p a(k)R(-k, -i) = -R(0, -i) \quad (25)$$

$$E_{\min}' = R(0,0) + \sum_{k=1}^p a(k)R(0,k) \quad (26)$$

Nonstationary processes are not ergodic, so the ensemble average is not equivalent to the time average. However, for locally stationary processes (speech can be considered locally stationary), it is reasonable to estimate the autocorrelation function with respect to a point in time as a short-time average. We estimate $R(-k,-i)$ by the covariance function in (19) as with the autocorrelation. This results in our previous equation (17) (Makhoul 1975).

LPC Speech Model

In our discussion of the general LPC model, we determined how to find the predictor coefficients. We did not determine, however, what type of excitation signal we should use as input to our time-varying filter or how many predictor coefficients will be required for our filter. We will now describe our speech model more concretely.

In a person, the excitation pressure (built up in the lungs) is acted upon or "filtered" by the shape and movement of the channel it flows through (see Figure 5).

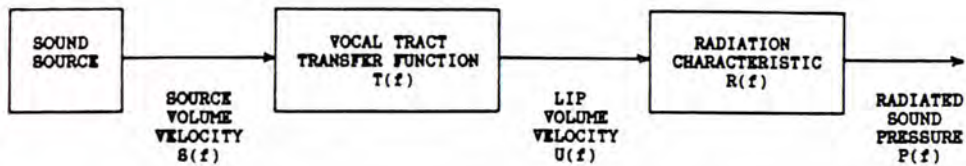


Figure 5. Speech production block diagram (Klatt 1980).

The frequency domain representation of the excitation multiplied by the transfer function of the filter gives the representation of the speech output in the frequency domain shown in Figure 6.

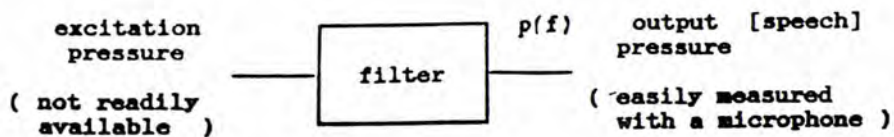


Figure 6. Basic speech production model.

It is difficult to separate the excitation source from the filter when analyzing speech because the only signal that is readily available is the speech itself. The excitation signal is not readily available. One method to simulate speech is to evaluate the frequency characteristics of the speech $[p(f)]$ and create a filter with transfer function $p(f)$. Then if a signal with a frequency spectrum equal to one is used as the input to the filter, the output will be $p(f)*1 = p(f)$. This is one way of looking at LPC speech synthesis. Both the delta function and normalized white noise have frequency spectrums equal to one for all f . Generally, white noise is used as the input for unvoiced sounds, such as /s/, and a train of delta functions is used as the input for voiced sounds, such as /e/. We will examine the excitation signal much more closely in the next chapter.

Filter Designation of the Speech Production System

In order to implement a filter for $p(f)$, it is necessary to have an idea of what we can expect the transfer function to be. The best way to see what type of transfer function is needed is to determine what the contribution by each part of the physical

speech production system will be (even though it is not necessary to determine these contributions separately in practice). The important contributions are from the glottal volume flow, the vocal tract, and lip radiation (Atal and Hanauer 1971).

Glottal Flow. The z transform of the glottal volume flow can be modeled by two poles. This is a good approximation during a single pitch period, which is all that is necessary since the filter coefficients are generally updated at least once every pitch period.

Vocal Tract Filter. The simple model of the vocal tract represents it as a linear filter ($T(f)$ in Figure 5). It can be shown that the zeros of the filter transfer function, when it has any, lie within the unit circle in the z plane. This allows each zero to be approximated by multiple poles in the transfer function. In addition, the location of the poles turns out to be much more important to how a person hears speech than the location of the zeros. So an all-pole filter can be used to model the vocal tract very effectively.

Radiation. The radiation of sound from the mouth (R(f) in Figure 5) is represented as a zero at $z=1$ in the z plane, but that zero along with a pole from the glottal flow or vocal tract transfer function

$$F(z) = \frac{(1-z^{-1})}{(1-az^{-1})} \quad (27)$$

can be approximated as a single pole. If we rewrite $F(z)$, we can see what the pole must be equal to.

$$F(z) = \frac{1}{\frac{(1-az^{-1})}{1-z^{-1}}} \quad (28)$$

By using long division, we can find the pole.

$$p(z) = 1-z^{-1} \left| \frac{1+(1-a)z^{-1}}{1-az^{-1}} - \frac{(1-a)z^{-1}}{1-z^{-1}} \right. \quad (29)$$

Therefore, $F(z)$ is equal to the pole minus the error term.

$$F(z) = \frac{1}{(1+(1-a)z^{-1})} - \frac{(1-a)z^{-1}}{(1-az^{-1})(1+(1-a)z^{-1})} \quad (30)$$

This can be combined again to show:

$$F(z) = \frac{1-az^{-1}-(1-a)z^{-2}}{(1-az^{-1})(1+(1-a)z^{-1})} \quad (31)$$

And the original form can be obtained.

$$F(z) = \frac{(1-z^{-1})(1+(1-a)z^{-1})}{(1-az^{-1})(1+(1-a)z^{-1})} = \frac{1-z^{-1}}{1-az^{-1}} \quad (32)$$

The contribution of this error to the transfer function is very small if one of the poles is close to $z=1$. This is usually the case in practice (Atal and Hanauer 1971).

Determination of the Number of Poles Needed for the Vocal Tract Filter

So the overall speech production model can be represented as a single discrete linear all-pole

filter $p(f)$ which is constantly changing in time (Childers, Yea, and Krishnamurthy 1986). The number of poles required to represent any speech segment adequately is determined by (1) the number of resonances and antiresonances (poles and zeros) of the vocal tract in the frequency range of interest; (2) the nature of the glottal volume flow function; and (3) the effect of radiation. Once the number of poles needed is determined, the speech production system can be implemented by one recursive filter without determining the contributions of each part of the system separately.

Two poles are adequate to represent the excitation, and the effect of radiation is combined with one of the poles with only a small error. Considerably more poles are required to represent the vocal tract transfer function. Theoretical evaluation of the vocal tract as an acoustic tube of varying cross-sectional area, as well as experiments simulating the vocal tract with various numbers of poles, reveals that about ten poles are required for a typical male voice. This makes the overall filter

order at least twelve for good LPC synthesis. We will examine the theoretical determination of the number of poles needed in the next section.

Acoustic Tube Model. In order to evaluate the number of poles needed for the vocal tract transfer function, we use the acoustic tube model. The vocal tract can be considered to be an acoustic tube of variable cross-sectional area. The relationship between the sound pressure P_g and the volume velocity U_g at the glottis and the same quantities P_l and U_l at the lips can be described by the ABCD matrix parameters as shown.

$$\begin{bmatrix} P_g \\ U_g \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} P_l \\ U_l \end{bmatrix} \quad (33)$$

The inverse Fourier transforms of these parameters have finite duration $t = 2l/c$, where l = length of the tube and c is the speed of sound. The function $S(x)$ shown in Figure 7 is the area of the vocal tract x distance away from the glottis. The

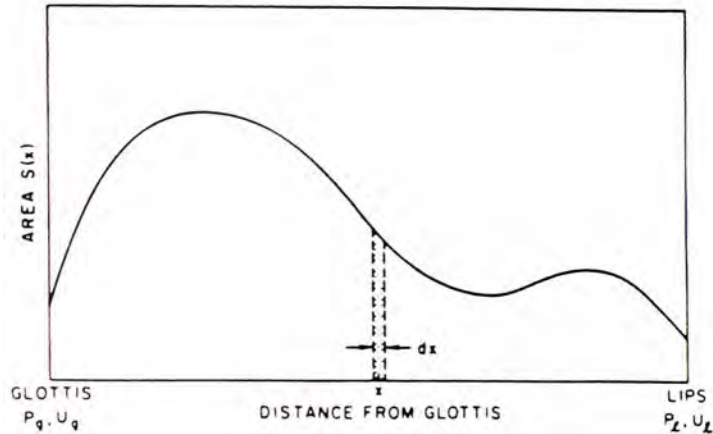


Figure 7. Vocal tract area function (Atal and Hanauer 1971).

ABCD matrix of the complete tube is given by the product of the ABCD matrices of n sections of tube each dx long ($l = n \cdot dx$). Each of the ABCD parameters for the entire tube can then be expressed in a power series of the form

$$\sum_{k=-n}^n d(k) e^{kGdx} \quad (34)$$

where G is the propagation constant for air. The ABCD parameters are therefore Fourier transforms of functions in time with duration $t = 2 \cdot n \cdot dx / c$. Let dx go to 0 and n go to infinity and then since $n \cdot dx = l$, $t = 2 \cdot l / c$.

So the filter must be able to model a function with duration t . Since the filter is made up of terms

which are the filter coefficients multiplied by z to negative powers up to $-p$, where p is the filter order, the maximum delay in the filter is $T \cdot p$, where T is the sampling period (each $1/z$ is a delay of T seconds). So $T \cdot p$ must be greater than or equal to $2 \cdot l/c$.

Most of the energy in the speech signal is below 5 kHz, so the signal is often sampled at 10 kHz. A typical number for the length of the vocal tract of a male is 17 cm. and the speed of sound in air is 340 m/s. This gives p greater than or equal to

$$\frac{2 \times l}{c \times T} = \frac{.34 \text{ m}}{.034 \text{ m}} = 10 \quad (35)$$

Again this makes the overall filter order 12 (Atal and Hanauer 1971).

Synthesis

Figure 8 is a block diagram of our speech synthesizer. The parameters in the figure that we have not yet determined are the pitch period (for voiced speech), the voiced-unvoiced parameter, and the rms value of the speech samples (G). The rms value of the speech samples is easily computed. Several

different procedures are available to determine if a speech segment is voiced and if it is to determine its pitch period. The voiced-unvoiced decision is often based on the fact that unvoiced speech has many more

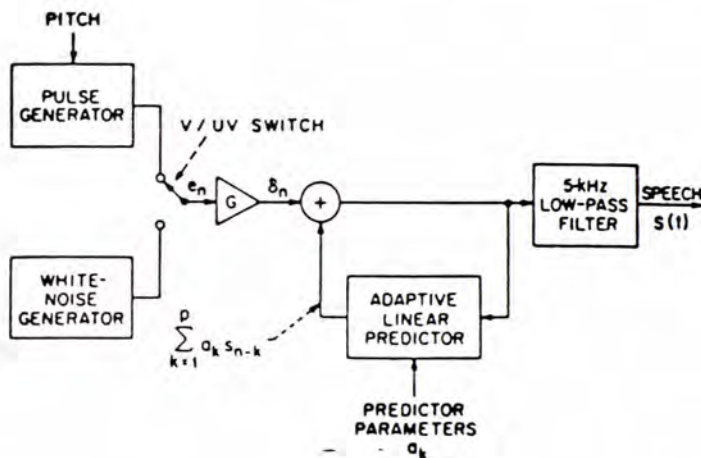


Figure 8. Block diagram of a speech synthesizer (Atal and Hanauer 1971).

zero crossings in a given time than voiced speech. The pitch period determination makes use of the fact that the fundamental frequency will be less than 1 kHz.

The pulse generator produces a pulse of unit amplitude at the beginning of each pitch period. The white noise generator produces uncorrelated uniformly distributed random samples with standard deviation

equal to one. The selection between the two generators is made by the voiced-unvoiced switch. The amplitude of the output signal is determined by the gain (G). The predicted value $s'(n)$ is added to the excitation to produce the synthetic speech. Then the signal goes through a low-pass filter which produces the continuous speech. We now have a complete model to synthesize speech. Next, we will examine altering the model to enhance the "naturalness" of the output speech.

CHAPTER 3

EXCITATION SIGNAL

We have seen how to obtain the linear predictor coefficients, the voiced-unvoiced decision, the rms value of the speech signal, and the fundamental frequency for our speech model. We also saw an example of one type of excitation, single impulse excitation. Now we will look at the excitation signal more closely for the reason that it has a large impact on the naturalness of the speech that will be obtained. First we will examine the ways we can obtain the excitation signal. Then we will look at the characteristics that an excitation signal should have. Finally, we will look at the seven different excitation signals to be used in our study. One of the seven is the excitation signal created by the author specifically for producing natural-sounding speech. The new excitation signal is very promising and it did well in our listening tests.

Methods for Constructing the Excitation Signal

In our LPC analysis we separate speech into two distinct parts. The first part is a filter determined by coefficients, and the second part is the excitation which is a residue signal. By using the residue signal as the excitation to the filter we could exactly reconstruct the original speech signal. Keeping the complete error signal, however, would result in either high bit rates for transmission or require extensive data storage capability. This destroys a major use of LPC which is compressing data. Three categories of methods used to create an excitation signal that approximates the residual signal, but still results in smaller storage requirements and data rates will be discussed (Naik 1984).

Residue Excitation

For residue excitation the excitation signal is created using the original residue signal. Some part of the frequency spectrum of the residue signal is used. Then an approximation of the full spectrum of the residue is generated and the excitation signal is created. Generating the spectrum is imprecise and this

results in less natural speech. Residue excitation and voice excitation do not depend on a voiced-unvoiced decision (Naik 1984).

Voice Excitation

This is similar to residue excitation except that some part of the original speech frequency spectrum is sent. Then the excitation signal is generated by flattening the spectrum that is received. Voice excitation produces speech which is even less natural than when residue excitation is used (Naik 1984).

Parametric Excitation

For parametric excitation, voiced and unvoiced excitations are generated with two distinct models. The parameters of these models include fundamental frequency, gain, and the voiced-unvoiced decision. The parameters are calculated continuously. This method allows parameters to change independently which causes parametric excitation to be useful for understanding speech synthesis. We will use only parametric models for our study. Next we will discuss the requirements that an excitation signal should satisfy.

Excitation Signal Requirements

Flat Spectrum

The predictor coefficients determine a filter with the same frequency envelope as the original speech. So the excitation signal must have a flat spectrum if the synthesized speech is to have the same spectrum as the original speech. Both the impulse function and random white noise have flat frequency spectrums. Other excitation signals have been suggested which have smoother shapes, but they often result in "bassy" speech because the lower frequencies are overemphasized (Naik 1984).

Broad Pulse Shape

LPC synthesis which uses parametric excitation models often produces speech with an unnatural "buzziness." The short duration of the pulses used in voiced excitation is considered one of the main reasons for this buzziness. The prime example is the single impulse excitation with which we developed the LPC model. The pulse is only one sample long, so after the excitation drops to zero, the synthesized speech decays quickly and remains at a low energy until the next pulse excites the filter. Therefore,

the output has large values at the excitation points and relatively small values in between. The excitation should ideally have a broader pulse shape to reduce buzziness, but not so broad that the spectrum loses its flatness, since this will result in bassy speech.

Association with Volume Velocity Waveform

The excitation model should take into account the actual volume velocity waveform at the glottis. This waveform results from the physical opening and closing of the glottis (Naik 1984). Making the excitation correspond to actual physical movement results in an increase in naturalness for the synthesized speech. The Electroglossograph (EGG) is very useful in determining these movements continuously. In fact, improved speech can be synthesized if the EGG signal as well as the speech signal is used in the analysis. However, for our parametric models we will assume a fixed ratio between the time to specific points in the volume velocity waveform and the overall pitch period.

Seven Different Parametric Excitation Signals

Unvoiced excitation is random noise in all of the models used in our study. This gives very natural sounding unvoiced speech segments. Therefore, the naturalness of the synthetic speech is primarily determined by the voiced segments. So only the voiced excitation will be described for each type of excitation signal.

To illustrate the different excitation signals, we will use three pitch periods of the sustained vowel /e/ as in hello. These three pitch periods of the speech waveform, 20.6 ms long, are shown in Figure

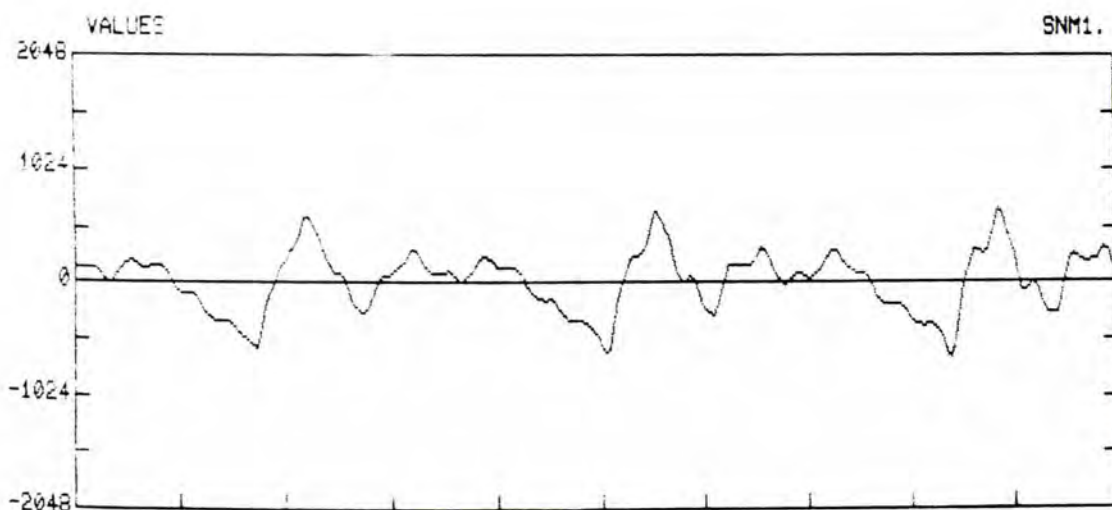


Figure 9. Three frames of recorded speech.

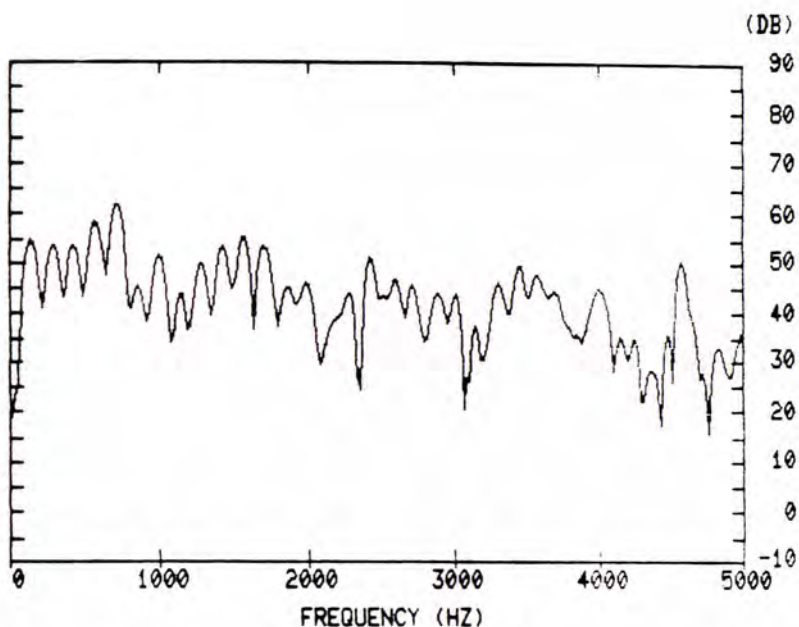


Figure 10. FFT of natural speech.

9. The fundamental frequency for this segment is $3/20.6E-3$ or approximately 150 Hz. This can also be seen in Figure 10, which is a plot of the FFT of the speech signal. Figure 10 also shows the formant frequencies at approximately 500, 1500, 2500, 3500, and 4500 Hz.

Single Impulse Excitation

This is the excitation with which we developed our model. It consists of a single pulse of only one point at the beginning of each pitch period as shown in Figure 11. This excitation's strongest point is

its simplicity. It is able to produce very intelligible speech, even though the synthetic speech waveform does not resemble the original speech. The synthesized speech is shown in Figure 12. The short duration of the excitation results in the filter response decaying to small values very quickly after each pulse. This is believed to be one of the reasons for the buzziness problem of speech synthesized with single impulse excitation. The FFT of the single impulse excitation is shown in Figure 13. Remembering that the ideal spectrum should be flat, Figure 13 shows a very ideal frequency spectrum. So this

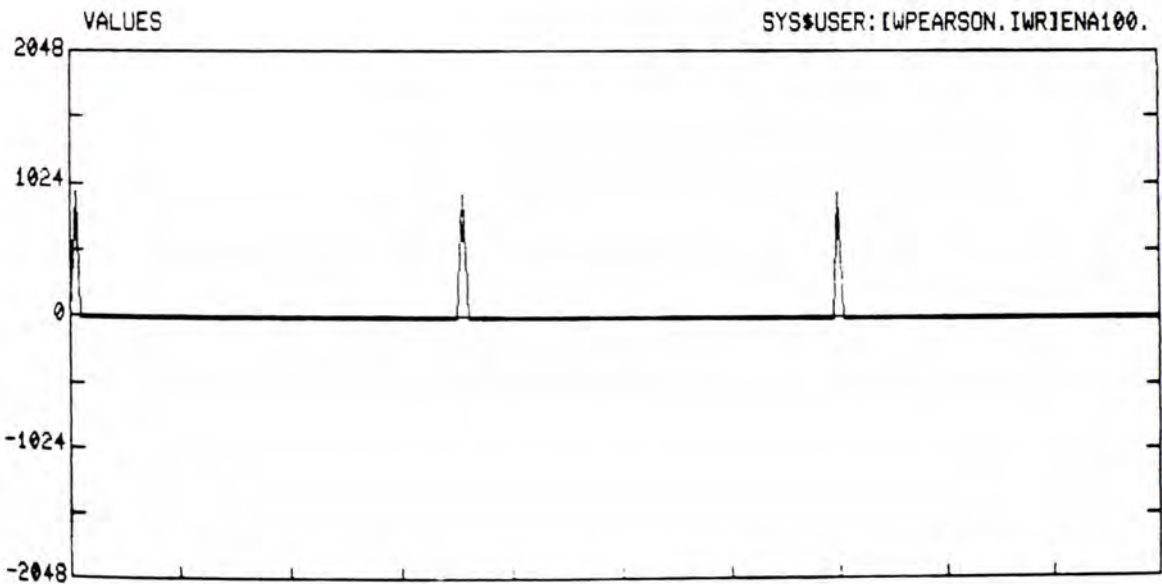


Figure 11. Single impulse excitation.

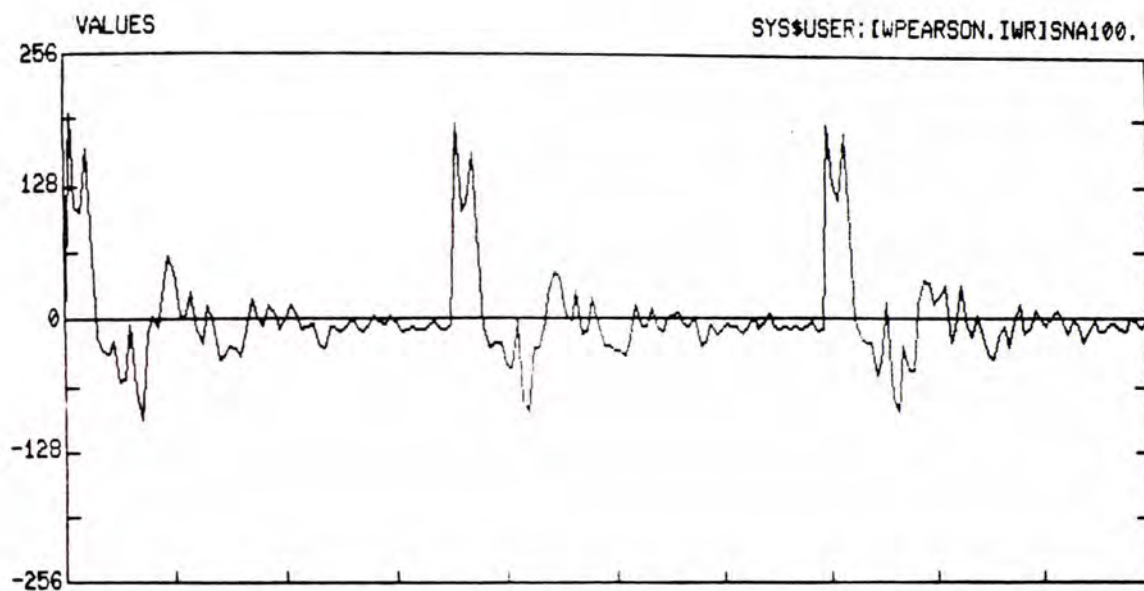


Figure 12. Speech synthesized using single impulse excitation.

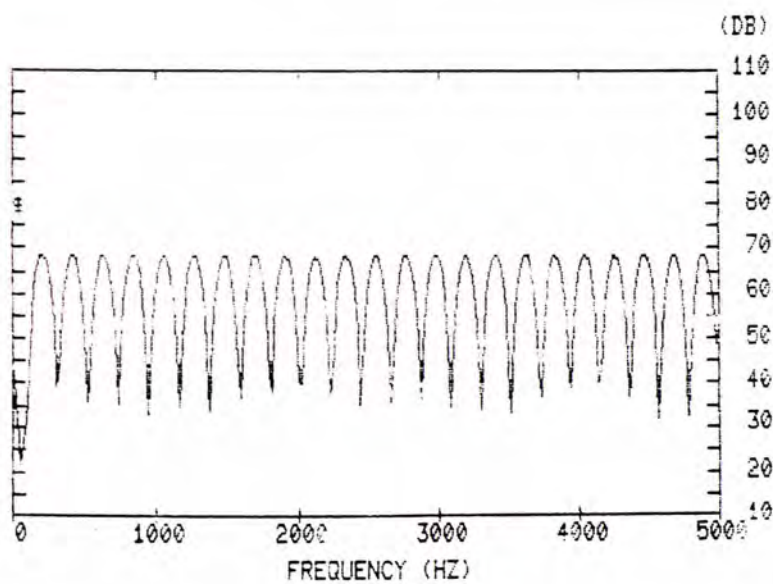


Figure 13. FFT of single impulse excitation.

excitation has a good frequency characteristic but a poor time wave characteristic (it is not broad enough). Although this excitation is useful in obtaining intelligibility, it does not produce very natural-sounding speech.

Double Impulse Excitation

This excitation is very similar to single impulse excitation except that it adds a second, smaller impulse which corresponds to the beginning of the glottal opening phase. The original impulse corresponds to glottal closing. The second impulse is placed a fixed percentage of the way into the pitch period as shown in Figure 14. Double impulse excitation corresponds a little better to the actual speech excitation than single impulse excitation. The synthesized speech also has less problem with decaying to small values, since the filter is excited twice within each pitch period (see Figure 15). However, double impulse excitation still retains the buzziness of single impulse excitation. Figure 16 shows that the spectrum of the double impulse excitation is nearly as good as that of the single impulse excitation.

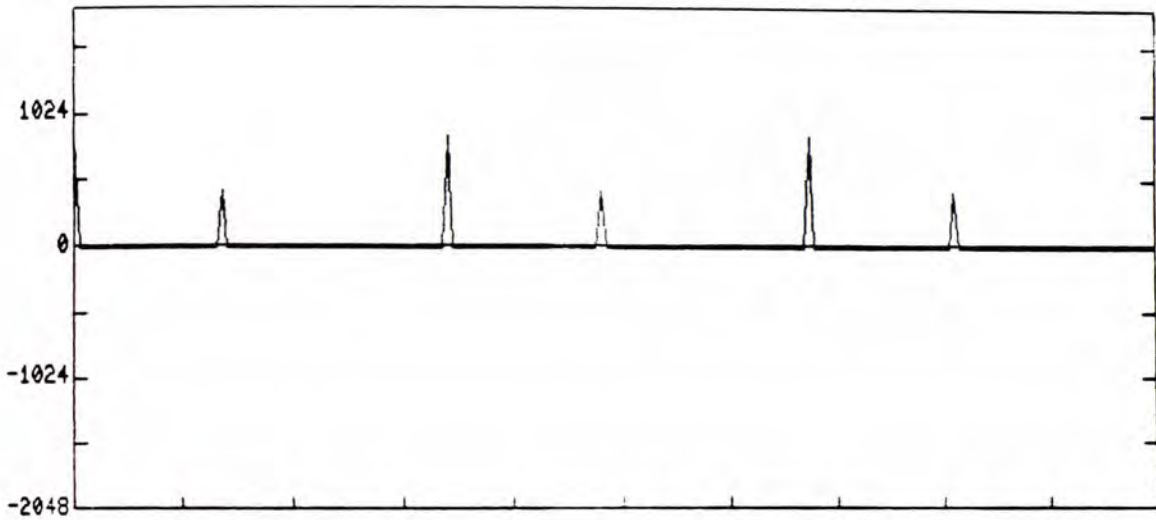


Figure 14. Double impulse excitation.

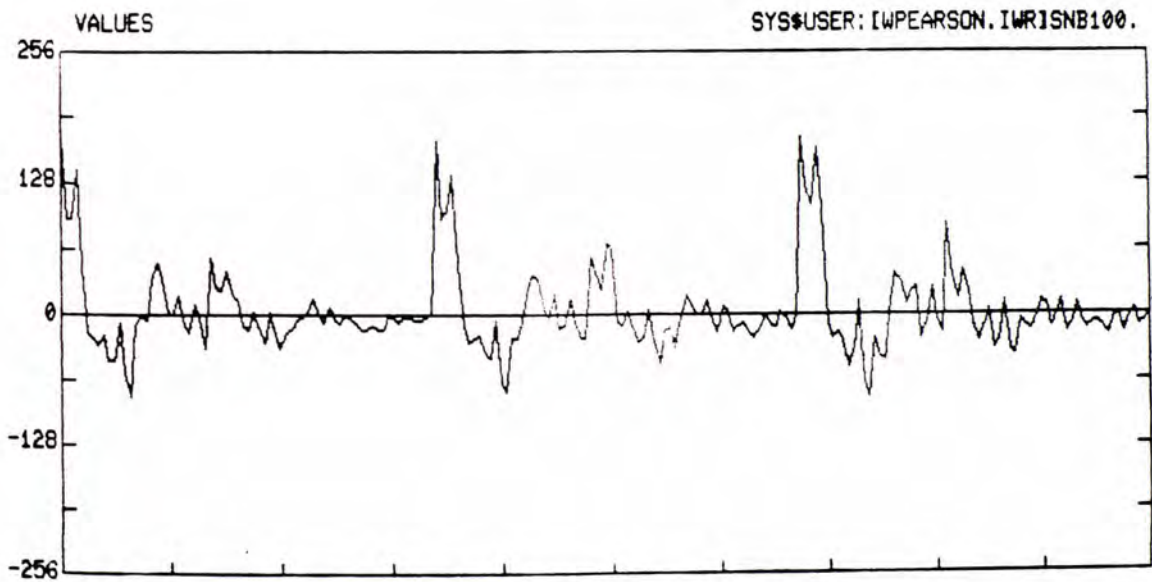


Figure 15. Speech synthesized using double impulse excitation.

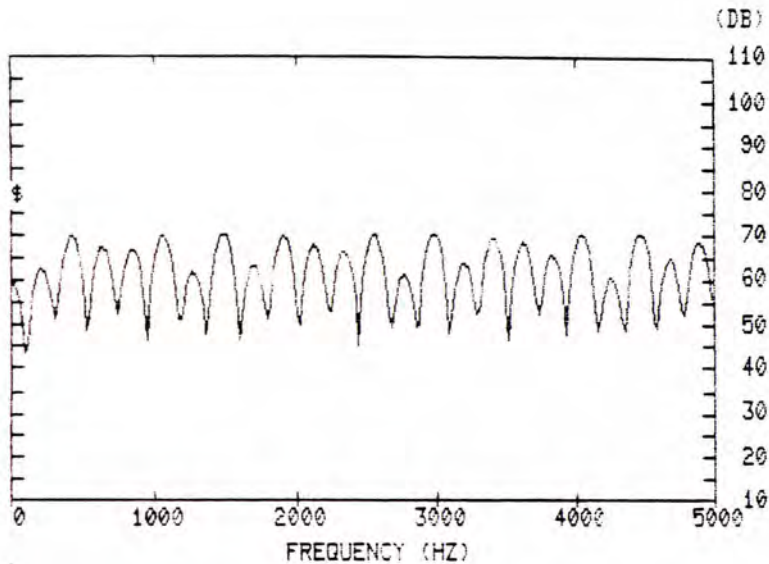


Figure 16. FFT of double impulse excitation.

Triple Impulse Excitation

This excitation is similar to the first two except that it adds a third impulse. So triple impulse excitation has pulses at the beginning and ending of the opening phase as well as at closing (Wu 1985). The excitation waveform is shown in Figure 17. The three pulses excite the filter more uniformly and all the major points of excitation in actual speech production. This results in less decaying of the filter response as can be seen in Figure 18. However,

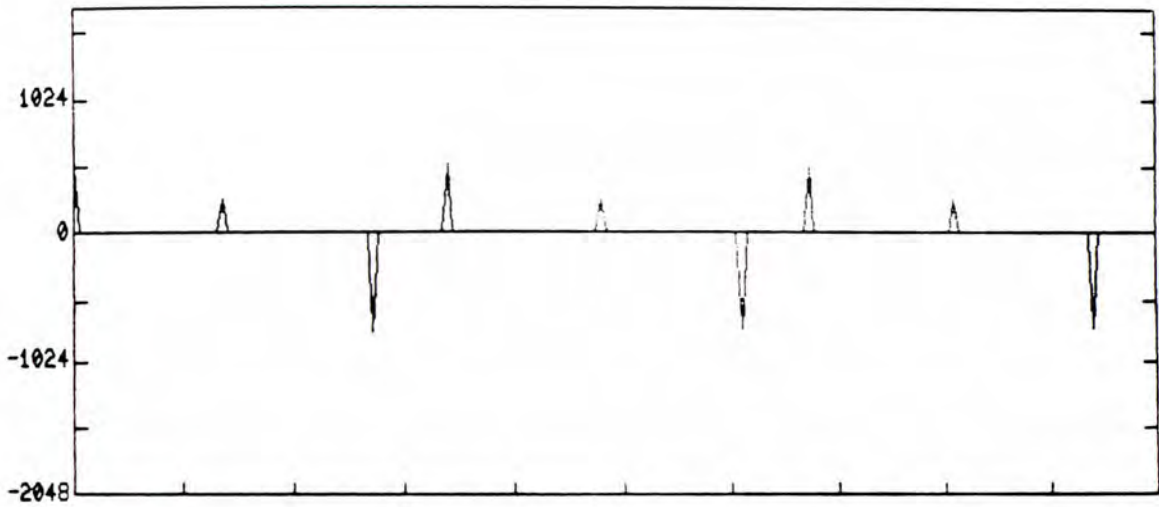


Figure 17. Triple impulse excitation.

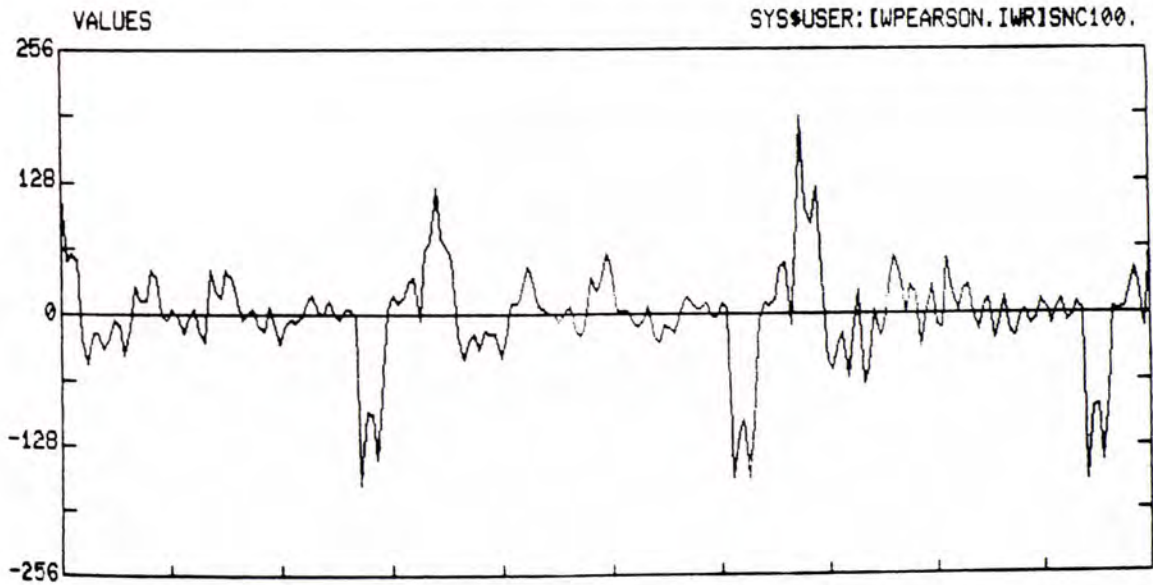


Figure 18. Speech synthesized using triple impulse excitation.

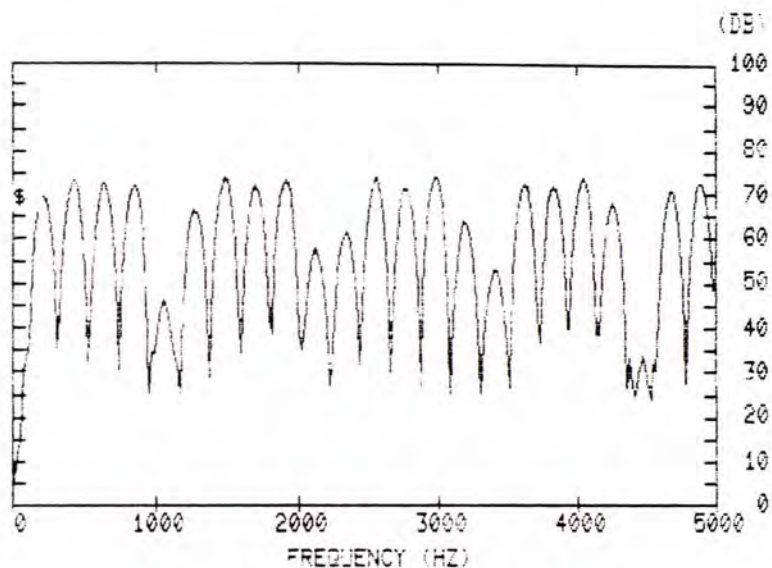


Figure 19. FFT of triple impulse excitation.

the synthesized speech still bears little resemblance to the original even though it also has a good frequency characteristic (see Figure 19).

First Derivative Of Fant's Excitation

The model of glottal volume velocity flow adopted by Fant is illustrated in Figure 20. The volume velocity is a measure of the flow of air through the glottis which is the excitation for natural speech. The equation for the rising branch is:

$$U = .5U_0[1 - \cos W_g(t - T_1)] \quad (36)$$

U_0 is the peak flow, W_g is the glottal frequency, and T_1 is the starting point for the pulse. The equation for the falling branch is:

$$U = U_0 [K \cos W_g (t - T_2) - K + 1] \quad (37)$$

K is the asymmetry factor. For $K = .5$, the falling branch is symmetrical to the rising branch. As K increases, the falling branch becomes shorter and steeper (Fant 1979).

For LPC synthesis, the first derivative of Fant's model is taken to enhance the spectral flatness (Wu 1985). The resulting excitation is shown in Figure 21. The equations used to calculate the excitation signal in discrete form are:

$$e(n) = .5E_0 \sin(W_g nT_s), \quad 0 \leq nT_s \leq T_1 \quad (38)$$

$$e(n) = -E_0 K \sin(W_g (nT_s - T_2)), \quad T_1 \leq nT_s \leq T_2 \quad (39)$$

$$e(n) = 0, \quad T_2 \leq nT \leq T \quad (40)$$

This excitation is very close to the excitation that produces human speech. Notice that (see Figure 23) the frequency spectrum overemphasizes the lower frequencies. The resulting synthetic speech, shown in Figure 22, looks very similar to the original speech. It is not surprising then that it also sounds very close to the original speech. More specifically, it sounds more natural (without buzziness or a tinny quality) than speech synthesized using impulses. This

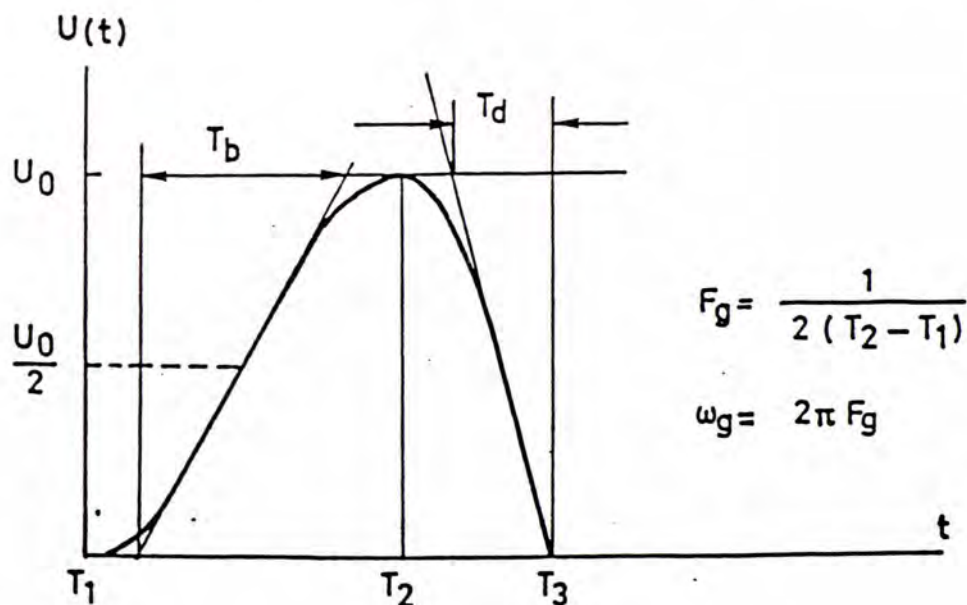


Figure 20. Fant's glottal flow model.

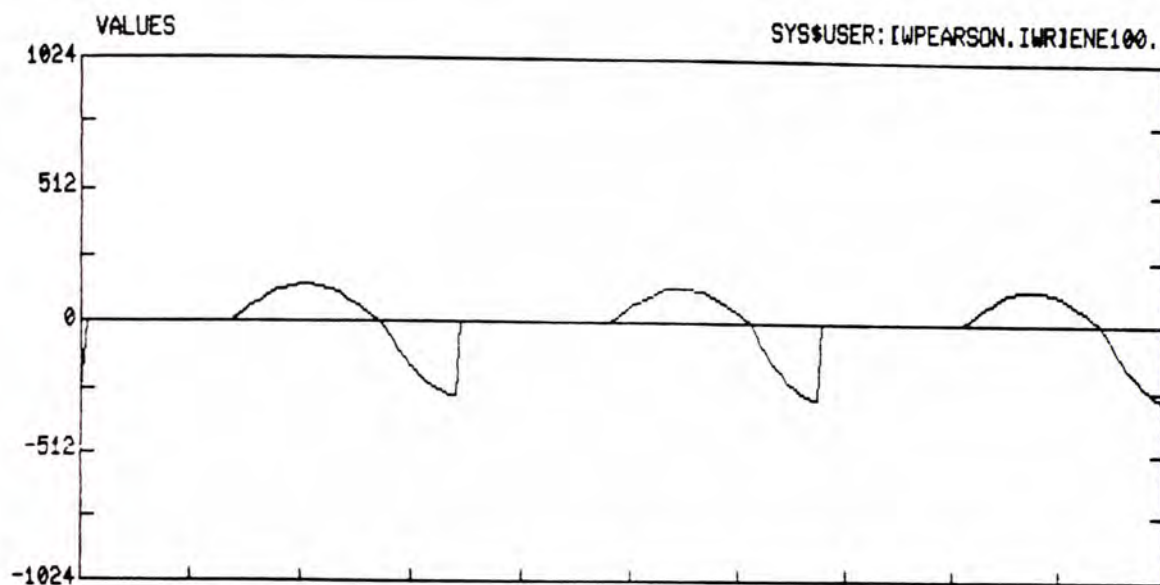


Figure 21. First derivative of Fant's excitation.

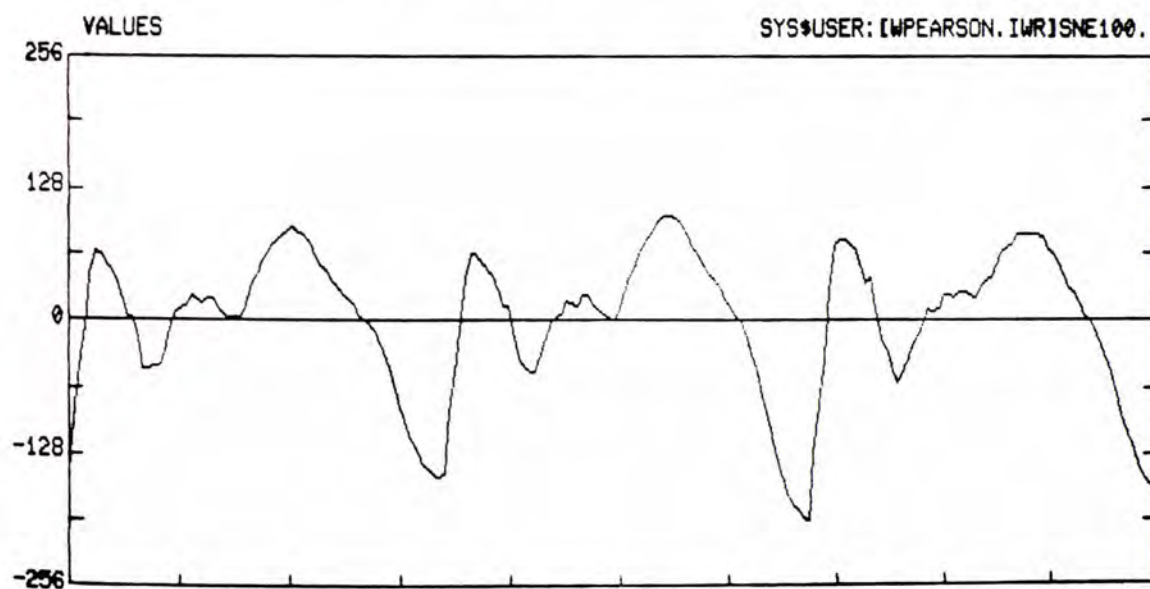


Figure 22. Speech synthesized using the first derivative of Fant's excitation.

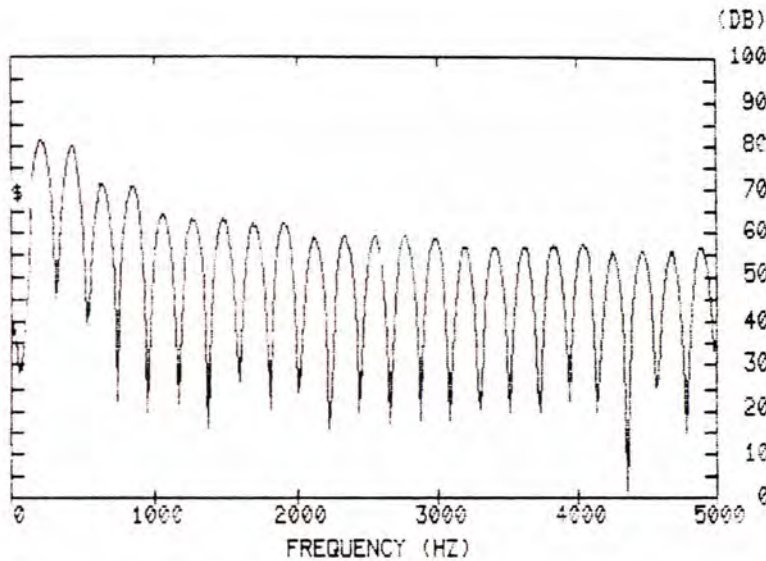


Figure 23. FFT of the derivative of Fant's excitation.

shows that some compromise between an ideal frequency characteristic and an ideal pulse shape may be necessary to produce speech that is both natural-sounding and intelligible.

LF Excitation

A minor disadvantage of using the first derivative of Fant's excitation is that it has a discontinuity at the flow peak (the large negative peak of the excitation) which adds a secondary weak excitation. Fant proposed a modified four parameter

model of glottal flow which is referred to as the LF model (Fant, Liljencrants, and Lin 1985). The fourth parameter is the time constant of an exponential recovery from the flow peak. The LF excitation is shown in Figure 24. The LF excitation is the same as the first derivative of Fant's excitation before the flow peak. The equation used to calculate the excitation after the flow peak is:

$$e(n) = -E_e [e^{(-b(nT-T_2))} - e^{(-b(T-T_2))}] ,$$

$$T_2 \leq nT \leq T \quad (41)$$

Like the first derivative of Fant's excitation, this excitation is close to the excitation that produces

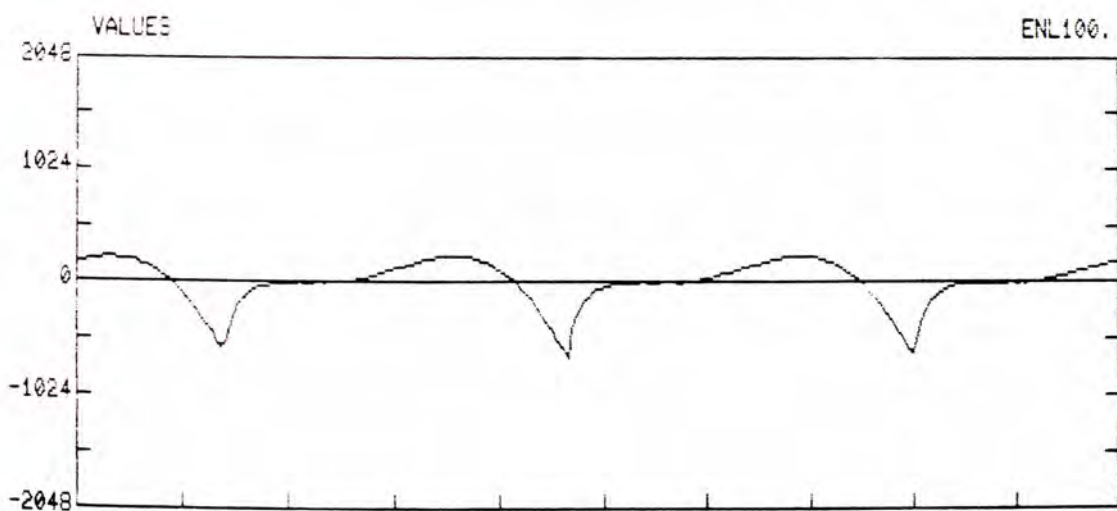


Figure 24. LF excitation.

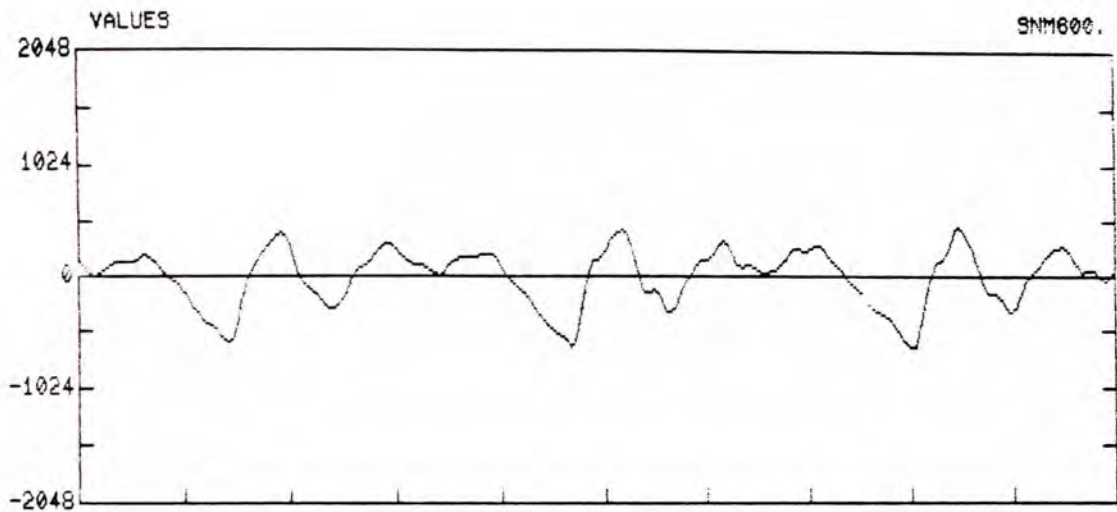


Figure 25. Speech synthesized using LF excitation.

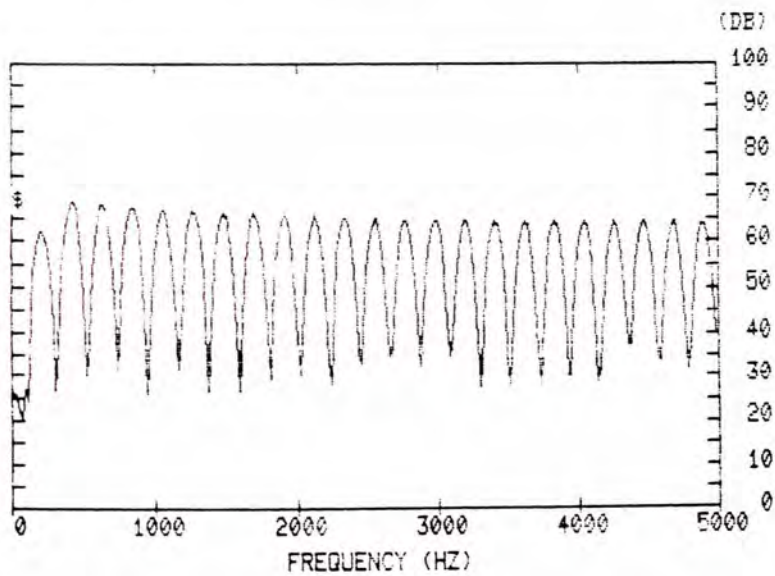


Figure 26. FFT of LF excitation.

human speech. But it has a flatter spectrum (see Figure 26). The resulting synthetic speech is shown in Figure 25. The quality and naturalness of this speech are very good.

Hilbert Transform

The Hilbert transform excitation most closely resembles the impulse excitations. It is formed by adjoining negative and positive pulses which represent the end of the opening phase and the beginning of the closing phase respectively (see Figure 27). The pulses do not start and end abruptly, however, as in the impulse excitation models. There is an exponential rise to the first pulse and an exponential decay from the adjacent pulse. The frequency spectrum, shown in Figure 29, is very similar to that of the single impulse excitation. But the resulting synthetic speech, shown in Figure 28, looks considerably better than the speech synthesized using single impulse excitation. The waveform does not decay to small values as rapidly. However, the excitation still results in speech which is buzzy - like speech synthesized using impulse excitation.

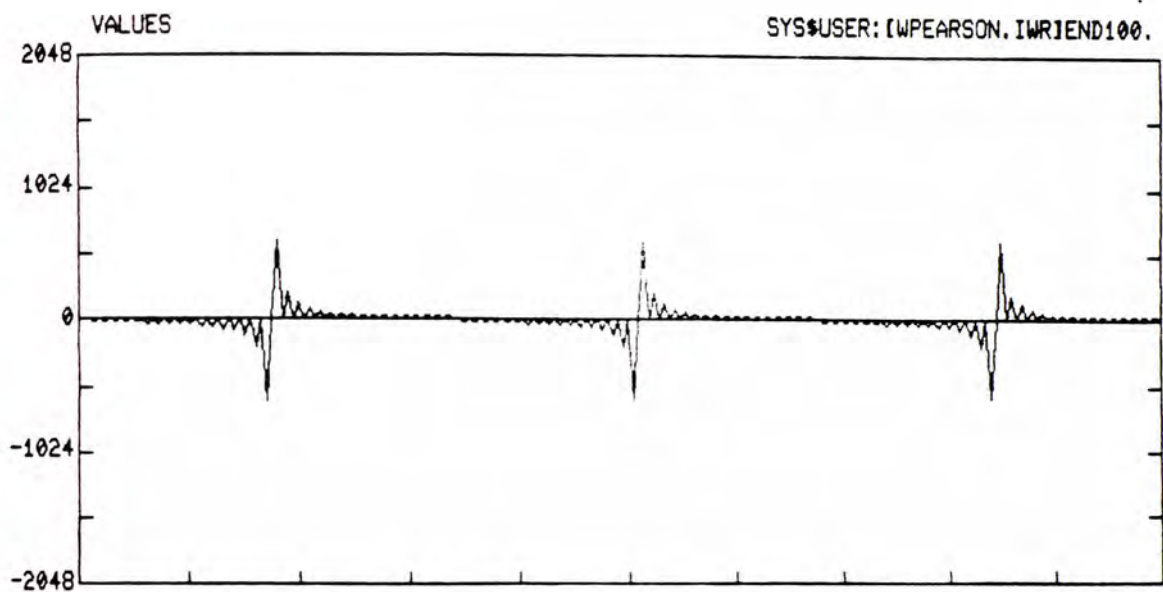


Figure 27. Hilbert transform excitation.

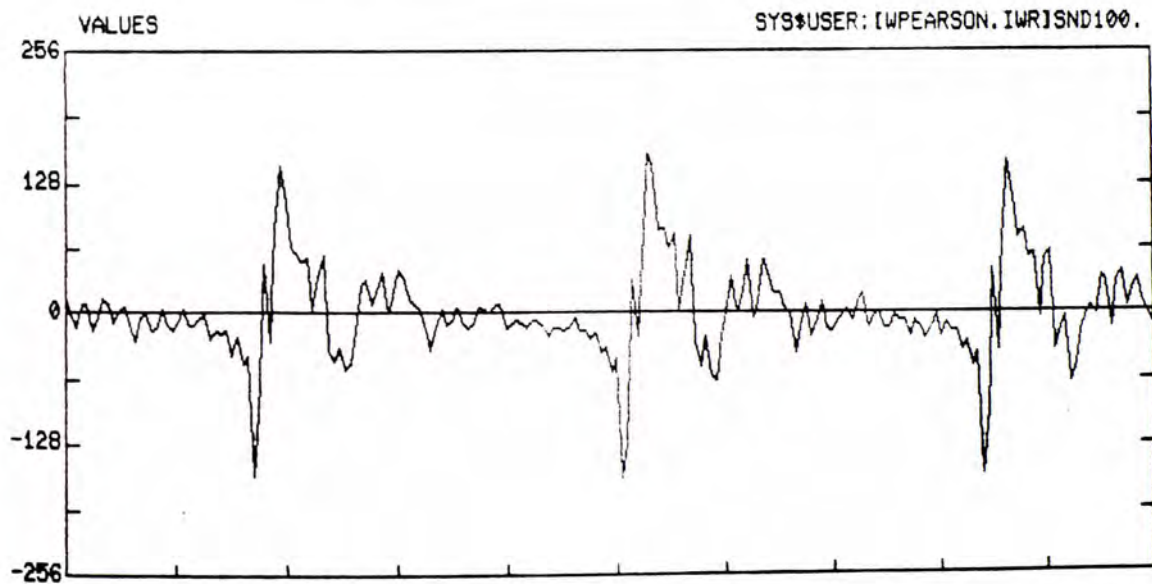


Figure 28. Speech synthesized using the Hilbert transform excitation.

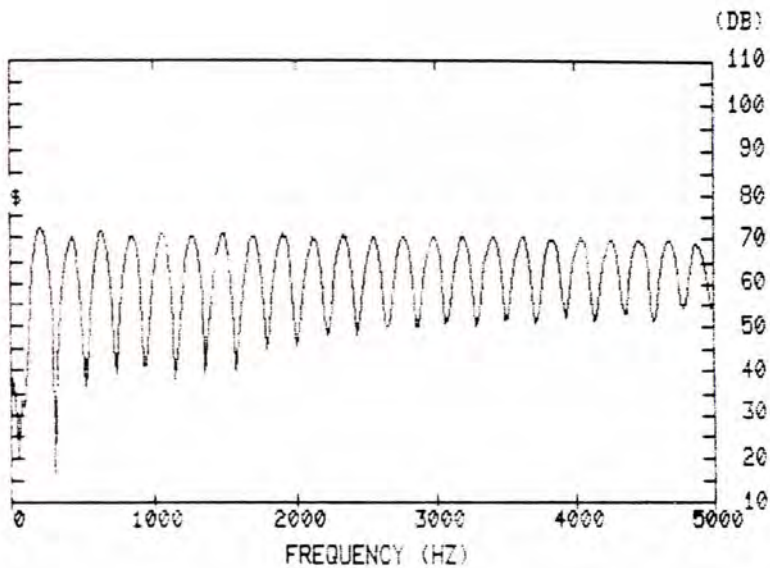


Figure 29. FFT of the Hilbert transform excitation.

LF Impulse Excitation

This is a new excitation created by the author in an attempt to retain the broad pulse shape and naturalness of LF excitation along with the clarity of impulse excitation. The smooth pulse shape of LF excitation results in a loss of clarity in a speech segment. The synthesized speech loses its crispness which results in a loss of intelligibility. Impulse

excitation results in a quickly decaying speech waveform with a buzzy quality in the speech segment. This buzziness sounds unnatural to the listener.

LF impulse excitation is created by using the LF pulse as an envelope for impulses. The impulses are spaced more closely together at the end of the pitch period (glottal closing) where the greater excitation occurs in natural speech. The LF impulse excitation is shown in Figure 30. The speech synthesized using LF impulse excitation is shown in Figure 31. The synthetic speech can be seen to have a broader pulse shape than speech synthesized using single impulse excitation. In fact, the synthesized speech shows a close resemblance to the shape of the original speech waveform. Figure 32 shows the FFT of this excitation. The figure shows the spectral flatness of LF impulse excitation. The resulting speech sounds smoother and more natural than speech synthesized using other impulse excitations. But it retains the crispness of speech synthesized using impulse excitation, which is lacking in speech synthesized using LF excitation.

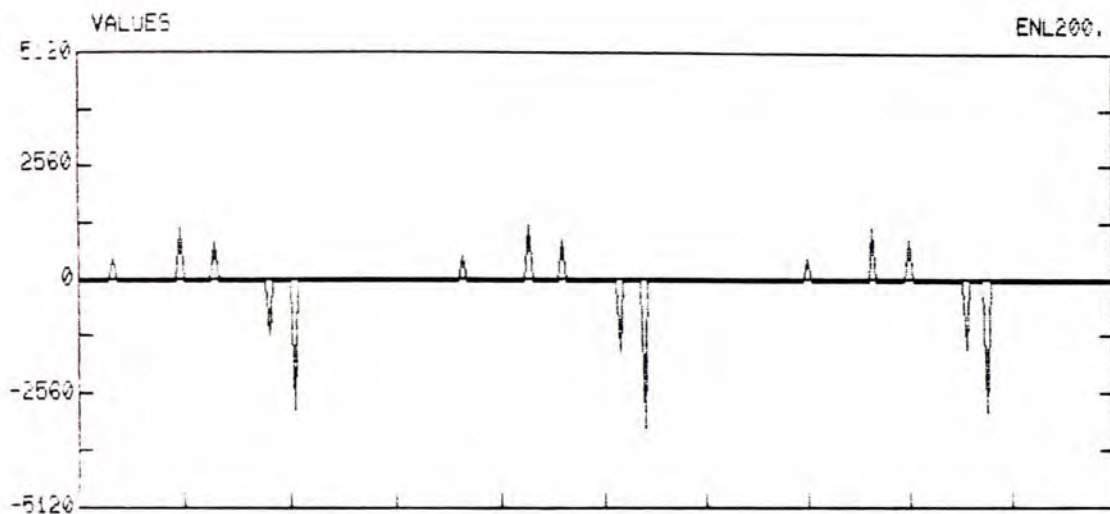


Figure 30. LF impulse excitation.

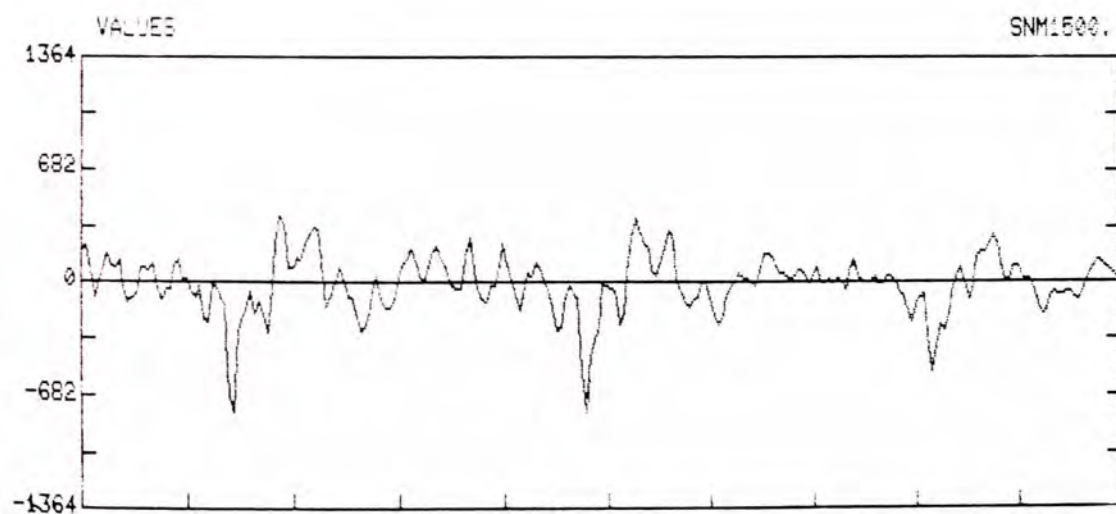


Figure 31. Speech synthesized using LF impulse excitation.

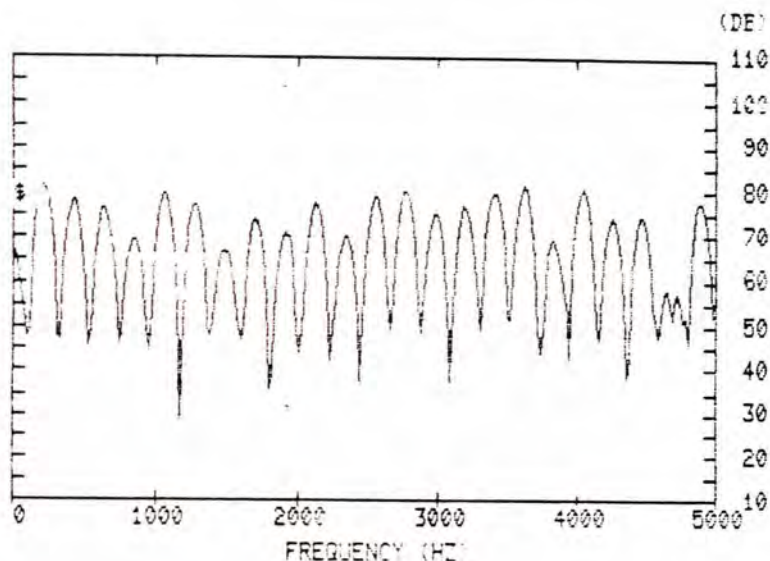


Figure 32. FFT of LF impulse excitation.

Conclusion

One promising new method for producing the excitation signal that we did not examine, because of its requirements for increased storage and completely different analysis techniques, is multipulse excitation (Jain 1983). This method does not make a distinction between voiced and unvoiced speech. It places pulses at key points (for example where the prediction error is large) for both voiced and unvoiced speech. This results in smoother speech since actual speech often has a mixture of voiced and

unvoiced excitation. This method does require increased bit rate and storage requirements (on the order of 50%), but it shows good potential and should be examined further.

We have seen the way that the different parametric excitation signals were derived, taking into account the requirements for an excitation signal. We have seen how the corresponding synthesized speech differs for different excitations. The most important factor, however, is how the speech actually sounds. This will be examined in the next chapter.

CHAPTER 4

TESTING FOR NATURALNESS

This chapter will examine the means to be used to test speech for naturalness. First, a brief description of the hardware necessary for our experiments is presented (a more detailed description is in Appendix D). Then we will describe the software involved in processing data, analyzing speech, and synthesizing speech. Our objective is to determine the type of excitation signal that can produce the most natural-sounding speech. However, we will also examine some of the other factors which can be important to naturalness for thoroughness. Next will be a description of the method which is used to compare speech synthesized using different excitation signals (forced pair trials). A report of the results of our testing will be followed by a conclusion about which type of excitation signal produces the most natural-sounding speech.

In order to run our experiments we set up an audio system connected to a Vax computer through an A/D - D/A system. In addition to installing the components, we designed and built the four linear phase low pass filters required. Next, we adapted the analysis and synthesis programs to run on our Vax system. Then we further modified them to interface with our Interactive Laboratory System Software (ILS). We created many new excitation signals in order to produce more natural-sounding speech. We experimented with our new excitation signals and the best of the classic excitation signals in use to evaluate their theoretical and subjective merits. We experimented with the analysis and synthesis parameters to determine which combination of parameters produced the best speech with each excitation signal. Then we used the best excitation signals in subjective speech trials with two classes of students. Our evaluation of the speech trials resulted in our conclusions about the type of excitation signal which produces the best quality speech.

Hardware

We will briefly describe the hardware required for our testing. More details are shown in Appendix D. The natural speech is sensed with a microphone and then lowpass filtered before being sent to the A/D to be digitized. The digitized data is stored on a Vax 11/750 computer. After the synthesized signal has been created (in digitized form) it is sent to the D/A and lowpass filtered. This signal is then amplified and played through speakers or headphones.

Software

Interactive Laboratory System Software (ILS) is a software package written by Signal Technology Inc. The basic speech processing programs used in our experiments was originally written at the University of Florida and modified and expanded at the University of Central Florida by the author. First the programs were adapted for use on our Vax computer. They were also modified to make them compatible with ILS software. Many new subroutines were added and the capability for using ten new excitation signals was

created. This made our study of many different excitation signals possible. Figure 33 shows the flow of data through the software involved.

ILS

ILS is a user-level software system with broad applications and capabilities for interactive digital signal processing and speech processing. ILS performs numerous functions in a variety of areas including: signal display and editing, data manipulation, spectral analysis, digital filtering and A/D - D/A data conversion.

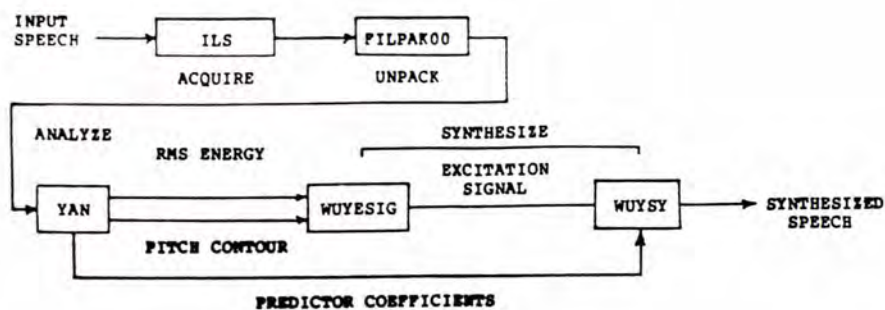


Figure 33. Flow of data through software in speech synthesis.

Filpak00

Since ILS stores data in packed form with a header, it is necessary to unpack the data before it can be used with most programs. This is done by filpak00 which places the unpacked information in a data file.

Yan

Yan is the analysis program used in our system. From the digitized speech signal it extracts the voiced/unvoiced determination, the pitch contour, the rms energy of the signal, and the predictor coefficients and stores them in data files.

Wuyesig

Wuyesig is the program that computes the excitation signal. Most of our programming effort involved Wuyesig. It takes the pitch contour and rms energy data and produces the desired type of excitation signal which is then stored in an ILS type file. A flowchart showing the operation of Wuyesig is shown in Figure 34. A complete listing of Wuyesig is included in Appendix E.

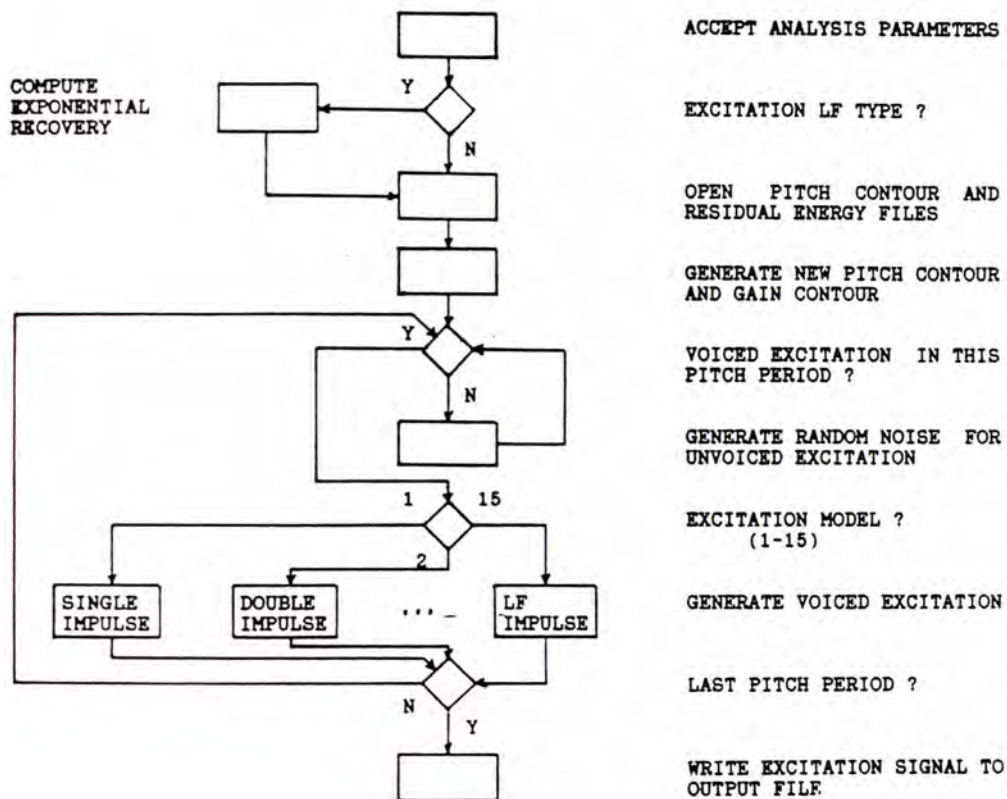


Figure 34. Flowchart for program Wuyesig.

Wuysy

Wuysy is the synthesis program that computes the synthetic speech from the excitation signal and the predictor coefficients. The synthesized speech is then stored in an ILS type file.

Other Factors Important to Naturalness

Although the purpose of this paper is to examine the effect of the excitation signal on naturalness, there are other factors which can be very important in determining the naturalness of synthesized speech. These other factors include the number of predictor coefficients, the sampling rate, the analysis method, the window type, the window length, the window increment, and whether pitch synchronous synthesis is used. This section will explain how we will minimize the effect of these other factors so that we can concentrate on the excitation signal.

Number of Predictor Coefficients

It has been well documented (Atal and Hanauer 1971; Pearson 1986) that the number of predictor coefficients used in LPC synthesis does not significantly affect the quality of the speech produced if at least a minimum number of coefficients (generally accepted as twelve) is used. In practical systems, it is usually necessary to use as few predictor coefficients as possible to reduce memory requirements and transmission rate. For our purposes,

however, we will use enough predictor coefficients (twelve) to assure that the synthesized speech is not degraded.

Bandwidth

Most of the energy in normal speech is contained in frequencies below 5 kHz. In practical systems, it is sometimes necessary to reduce the bandwidth of the signal to 3 kHz or less even though this degrades the speech quality. For our experiments, we will use a bandwidth of 5 kHz so that the speech is not degraded. This requires a sampling rate of 10 kHz.

Analysis Method

In Chapter 2, we discussed the two different analysis methods: autocorrelation and covariance. It is unclear which method is better for producing high quality speech. One method may not be superior for all situations. There is generally not a large difference in the speech produced using the two methods. For our purposes, however, we will compare the different excitation signals using each method to determine which method is best for each excitation signal.

Window Type

Since we may be examining a long speech segment, it is necessary to break the speech up into segments for analysis. This is done using the windows that were mentioned in Chapter 2. The choice of a particular window may make a large difference in the quality of the resulting speech. We will compare speech synthesized with four different windows for each excitation signal and determine which sounds the best with each excitation. The windows that we will use are the rectangular, hamming, hanning, and cosine windows.

Window Length

The best length for the analysis window involves a tradeoff between time domain and frequency domain resolution. A shorter window gives better time domain resolution and a larger window gives better frequency domain resolution (Childers and Wu 1986). The window must be long enough to contain one period of the lowest frequency in the signal in order to detect the fundamental frequency. For normal speech, the lowest fundamental frequency is about 100 Hz, so the window length should be at least 10 ms. The window should

not be much longer than 10 ms, though, or short time speech events will be degraded. For our tests, we will use a window length of 20 ms.

Window Increment

It is good practice to have the windows for adjacent frames overlap so that there is good continuity between frames. A shorter window increment causes successive windows to overlap more. But this also results in greater memory size and transmission rate requirements. We will use a window increment of 10 ms.

Pitch Synchronous Synthesis

In pitch synchronous synthesis, the synthetic speech is produced one pitch period at a time (for voiced speech) rather than producing a fixed length of speech each frame. This produces higher quality speech and will be used in our tests. It is also possible to analyze the speech synchronously. This can enhance the quality of the speech, but it results in a varying memory size and transmission rate requirement depending on the time-varying fundamental frequency of the speech. We will not use pitch synchronous analysis in our tests.

Comparison Method

Preliminary Evaluation

Before beginning our forced pair trials to determine the best excitation signal, we conducted a preliminary test to optimize the other factors which are important to naturalness. We found that there was no noticeable difference between speech synthesized using the autorrelation method and speech synthesized using the covariance method. We chose to use the autocorrelation method since it is always stable. We found that the hanning window clearly produced the most natural-sounding speech with every excitation signal. The hamming and cosine windows were almost as good and the rectangular window was considerably worse. Therefore we used the hanning window for our analysis.

In order to prevent the length of the listening test from being a strain on the listeners, we decided to drop two excitations from our test. Double impulse excitation and the first derivative of Fant's excitation were selected. Speech synthesized using double impulse excitation sounds very similar to speech synthesized using both single impulse

excitation and triple impulse excitation. Speech synthesized using the first derivative of Fant's excitation sounds very similar to speech synthesized using LF excitation except not quite as good. Dropping these two excitations allowed us to reduce the number of pairs from 168 to 90 and therefore reduce the test time from about 90 minutes to 45 minutes.

Forced Pair Trials

The instructions for our test and the listener evaluation form are shown in Appendix B. Each possible combination of the six speech segments (five synthesized and the original) was placed on audio tape twice. Each segment in the pair was placed first in the pair once and last in the pair once. This was to prevent bias due to the order in which the pair was presented. There were thirty pairs in each section. There were three sections - male voice, female voice and child voice. The order the pairs were presented in was random.

The sentence used was "We were away a year ago." It was chosen because it is a phonetically balanced sentence which contains voiced excitation. Voiced

excitation is considerably harder to synthesize well than unvoiced excitation. The listeners heard the original speech before beginning each section. For example, they heard the original male voice before the beginning of the trial section on male voice. They were asked to use the original speech as a standard for what "natural" speech should sound like for our test. So for comparison in our test, the original speech was defined as "natural."

As the tape was played, the listeners chose one speech segment from each pair as sounding more "natural" than the other. They circled the corresponding letter on the listener evaluation form. They were asked to choose a letter even if they were unsure which segment sounded better rather than leave any answers blank. This is where the term forced pair trials is derived. Our test was performed twice - once with ten beginning level speech students, and again with twelve graduate level speech students.

Results

Data Analysis

The programs used to analyze the results are listed in Appendix A. Each listener was checked for

consistency by checking the percentage of the time that the listener picked the same speech segment in a pair both times the pair was heard. A listener was considered consistent if the listener chose the same speech segment in a pair twice at least 75% of the time. Only the data from consistent listeners was used in our analysis. There were seven consistent students from each group or a total of fourteen listeners.

We determined the percentage of the time that each male speech segment was chosen. This was repeated for female speech and child's speech and all speech. The original speech was selected 98% of the time. The results for speech synthesized using each

TABLE 1
SPEECH EXCITATION RANKINGS

		<u>MALE</u>	<u>FEMALE</u>	<u>CHILD</u>	<u>TOTAL</u>	<u>RANK</u>
SINGLE IMPULSE	\bar{X}	50%	42%	40%	44%	4
TRIPLE IMPULSE	\bar{X}	38%	27%	30%	32%	5
HILBERT TRANSFORM	\bar{X}	50%	46%	49%	48%	3
LF	\bar{X}	63%	71%	73%	69%	1
LF IMPULSE	\bar{X}	49%	67%	59%	58%	2

excitation is shown in Table 1. Speech synthesized using LF excitation was ranked number 1. Speech synthesized using LF impulse excitation was ranked number 2. Speech synthesized using the Hilbert transform excitation and the single impulse excitation were ranked number 3 and 4. And speech synthesized using triple impulse excitation was ranked number 5.

The results show that LF excitation is probably the best excitation to use if the only concern is naturalness. This makes sense because impulse excitation generally results in a quickly decaying speech waveform. This causes the synthesized speech to be buzzy which results in a loss of naturalness. Since LF excitation is a continuous excitation (rather than being constructed of impulses), speech synthesized with it does not have this problem with buzziness. Speech synthesized with LF excitation sounds very smooth and natural.

Speech synthesized using LF excitation is, however, noticeably less intelligible than speech synthesized using any of the impulse excitations. This was noted by several students following the test and can be plainly heard. This loss of intelligibility results from the broad pulse shape in

LF excitation. This causes the synthesized speech to lose its crispness. Therefore, LF impulse excitation should be considered if a compromise between intelligibility and naturalness is needed. And some compromise between those two characteristics is generally needed in practical speech synthesis systems.

LF impulse excitation succeeds in producing speech which retains much of the broad pulse shape (which reduces buzziness) of LF excitation and all of the intelligibility of speech synthesized using impulse excitations. The speech retains the crispness of speech synthesized using impulse excitations. But, it sounds smoother and more natural than speech synthesized using other impulse excitations. This can be seen in the results shown in Table 1. LF impulse excitation was shown to be significantly superior to the other impulse excitations.

The results also showed that LF impulse excitation performed particularly well with the female voice. Therefore LF impulse excitation should also be considered if much of the speech to be synthesized is female speech. If intelligibility is the main

consideration then LF impulse excitation is probably a good choice (even though our study did not directly address intelligibility).

Finally, we conclude two very important points about producing high quality speech using LPC synthesis. First, some compromise is necessary between an ideal frequency characteristic and an ideal time waveform for the excitation signal. Second, LF impulse excitation is a good compromise between these two factors and therefore is an excellent choice for producing speech that is both very natural and highly intelligible.

APPENDIX A
LIST OF SOFTWARE

General Signal Processing Programs

1. ILS signal display, data manipulation,
 and A/D - D/A conversion.
2. Filpak00 unpack ILS data.

Speech Analysis-Synthesis Programs

1. Yan LPC parameter analysis.
2. Wuyesig computation of excitation signal.
3. Wuysy LPC synthesis.

Listener Data Evaluation

1. CC.BAS listener consistency check.
2. DA.BAS listener data check.
3. CDC.BAS combined listener data check.
4. IND.BAS input listener data and create
 data file.
5. INM.BAS input master pair list and
 create master file.

APPENDIX B
LISTENER EVALUATION TEST

INSTRUCTIONS

1. This is a listener evaluation session. It is designed to evaluate the performance of a speech synthesis scheme. The section has 3 different sections - one with a male voice, one with a female voice, and one with a child's voice. You will be presented with a series of sentence pairs - A and B. The sentence used is "We were away a year ago."
2. YOUR TASK is to judge WHICH sentence in each pair SOUNDS MORE 'NATURAL' than the other and circle the corresponding entry on the evaluation sheet. One of the sentences in each pair may or may not be the original speech.
3. At the beginning of each section you will hear the original speech repeated 3 times. This is to give you an idea of what 'NATURAL' speech should sound like. For comparison in our test, the original speech is defined as 'NATURAL'.
4. The presentation sequence is as follows:
The number of the sentence pair will be repeated.
The pair will be repeated twice (A,B) (A,B)
Then there will be a 4 second silence for marking your choice.
5. This sequence will be repeated for each pair. Please mark your choice only after each pair has been presented twice.

LISTENER EVALUATION FORM

Warmup

- 1 A B
- 2 A B
- 3 A B

Task I

- 1 A B
- 2 A B
- 3 A B
- 4 A B
- 5 A B
- 6 A B
- 7 A B
- 8 A B
- 9 A B
- 10 A B
- 11 A B
- 12 A B
- 13 A B
- 14 A B
- 15 A B
- 16 A B
- 17 A B
- 18 A B
- 19 A B
- 20 A B
- 21 A B
- 22 A B
- 23 A B
- 24 A B
- 25 A B
- 26 A B
- 27 A B
- 28 A B
- 29 A B
- 30 A B

Task II

- 1 A B
- 2 A B
- 3 A B
- 4 A B
- 5 A B
- 6 A B
- 7 A B
- 8 A B
- 9 A B
- 10 A B
- 11 A B
- 12 A B
- 13 A B
- 14 A B
- 15 A B
- 16 A B
- 17 A B
- 18 A B
- 19 A B
- 20 A B
- 21 A B
- 22 A B
- 23 A B
- 24 A B
- 25 A B
- 26 A B
- 27 A B
- 28 A B
- 29 A B
- 30 A B

Task III

- 1 A B
- 2 A B
- 3 A B
- 4 A B
- 5 A B
- 6 A B
- 7 A B
- 8 A B
- 9 A B
- 10 A B
- 11 A B
- 12 A B
- 13 A B
- 14 A B
- 15 A B
- 16 A B
- 17 A B
- 18 A B
- 19 A B
- 20 A B
- 21 A B
- 22 A B
- 23 A B
- 24 A B
- 25 A B
- 26 A B
- 27 A B
- 28 A B
- 29 A B
- 30 A B

NAME _____

APPENDIX C

USER'S GUIDE TO SPEECH SYNTHESIS PROGRAMS

USER'S GUIDE TO SPEECH SYNTHESIS PROGRAMS

This guide contains a step by step explanation of how to use the speech analysis and synthesis programs contained on the VAX1 at UCF under the directory [WPEARSON.IWR]. It also explains how to use the Interactive Laboratory Software (ILS) commands necessary to perform our synthesis. ILS is resident on the VAX1 operating system. The programs and their functions are listed below:

- ILS - signal display, data manipulation, and A/D - D/A conversion.
- Filpak00 - unpack ILS data.
- Yan - LPC parameter analysis.
- Wuyesig - computation of excitation signal.
- Wuysy - LPC synthesis.

These programs allow the user to produce Linear Predictive Coding (LPC) synthesized speech. Special emphasis is placed upon being able to choose from many different excitation signals for the synthesis model. Many other parameters, however, are selectable during both analysis and synthesis. This guide shows how to record a speech segment, analyze it, and produce synthetic speech from analysis parameters.

1. A sample speech synthesis session is shown at the end of this appendix. This explanation follows that example step-by-step. Page 100 shows how to record speech into an ILS file. Everything that appears on the screen is shown. At the far left are the commands or parameters that you enter.

first open a file:

```
$fil ansnw
```

sets the alphabetic characters at the beginning of your ILS file name to snw.

```
$ctx 1
```

sets the number of points in a frame of data to 1 (so 1 data point = 1 frame).

```
$fil crl,,,25600
```

creates the file snw1. with size 25600 frames, which is 101 blocks.

```
$ina sf10000
```

sets the sampling frequency to 10kHz. This is the clock rate for our A/D's and D/A's.

```
$fil
```

tells you information about the file currently in use.

```
$rec L1,1
```

sets the A/D system to record speech (up to 101

blocks worth, which is 2.6 seconds of speech) from channel 1.

\$dsp

displays the contents of your speech file graphically.

2. Page 101 shows how to unpack the data from its ILS file and begin the analysis:

\$r filpak00

begins the program to unpack your speech file. At the correct prompts you will enter your speech file name:

snwl.

the number of blocks it contains:

101

and the file you want the unpacked data to be written to.

andyul00.

\$r yan

begins the speech analysis program. You will enter your unpacked file name:

andyul00.

the number of points in the file:

25600

the window length to be used for the analysis:

200 (20 ms)

the amount the window is to be moved each iteration:

100 (10 ms)

the sampling frequency:

10000

the order of the linear predictor to be used:

12

the kind of window to be used:

(rectangular, hanning, hamming, or cosine)

1

the analysis method to be used:

(autocorrelation or covariance)

1

the program will respond with the number of frames being analyzed (which you will need to know later):

250

it will ask if you wish to compute the pitch period of each frame (yes or no):

1

the program will tell you the names of the files that contain the analysis data:

andyul00.lp

andyul00.rs

andyul00.pc

3. Page 103 shows how to produce the excitation:

\$r wuyesig

begins the production of the excitation signal.

You will be prompted for the number of frames in
the file:

250

the window increment:

100

the window length:

200

the file to contain the excitation signal
(should be an ILS type filename).

enll00.

the excitation model you choose to synthesize
with:

1

the file your pitch period values were stored
in:

andyul00.pc

the file your residual energy values were stored
in:

andyul00.rs

4. Page 104 shows how to produce the synthesized speech signal:

```
$r wuysy
```

```
begins production of the synthetic speech. You
will be prompted for the lp coefficient
filename:
```

```
andyul00.lp
```

```
the excitation filename:
```

```
enl100.
```

```
the synthesized output filename:
```

```
snw100.
```

```
the predictor order:
```

```
12
```

```
the number of samples per frame:
```

```
100
```

```
and the number of frames:
```

```
250
```

```
This completes our speech synthesis session. The
listen command (not shown in example) allows you
to hear your synthetic speech. First you must
set the default filename to snw100.:
```

```
$fil ansnw
```

```
$fil 100.
```

```
$lsn L1,25000
```

 this plays the first 25000 frames of your file.

Sample Speech Synthesis Session

```

$
$ fil ansnw
Alphabetic characters set to: SNW
$ ctx 1
CONTEXT =      1 POINTS/FRAME
$ fil crl,,,25600
SNW1.                                NOT SAMPLED DATA
    101 DK BLKS
PRIMARY FILE
$ ina sfl0000
SF = 10000                            SAMPLING FREQUENCY
$ fil
SNW1.                                SAMPLED DATA
    101 DK BLKS, 25600. FRAMES,      1 PT/FR
SAMPLE RATE = 10000 HZ
PRIMARY FILE
$ rec 11,1
RECORD INTO SNW1.
    1 CHANNEL(S) STARTING WITH CHANNEL 1
AT A SAMPLING FREQUENCY OF 10000. HZ
WITH A FILE HEADER OF 128 WORDS
AND FOR A TOTAL OF 2.6 SECONDS.

ENTER... A<RETURN> TO ABORT
ENTER... <RETURN> TO START A/D
TYPE ANY KEY TO TERMINATE A/D
->

===== DONE =====
$

```

\$ r filpak00

ENTER INPUT FILE NAME

snwl.

ENTER # OF BLOCKS

101

snwl.

101

ENTER OUTPUT FILE NAME

andyul00.

FORTRAN STOP

\$ r yan

**** SEQUENTIAL LINEAR PREDICTION ANALYSIS ****

THIS PROGRAM EVALUATES THE LP COEFFICIENTS, RES ENERGY
AND PITCH CONTOUR FOR A LONG SPEECH SEGMENT

THE V/UV/SILENCE DECISION IS MADE ON THE BASIS OF
AVERAGE ABS MAGNITUDE AND AUTOCORR PEAK THRESHOLD

50 MSEC. OF SILENCE IS REQUIRED IN THE BEGINNING
OF THE DATA RECORD FOR SILENCE DETERMINATION

INPUT FORMAT	:	BINARY INTEGER
MAX. WINDOW LENGTH	:	250 SAMPLES
MAX. PREDICTOR ORDER	:	40

** THERE IS NO LIMIT ON THE SIZE OF INPUT DATA ARRAY **

ENTER INPUT DATA FILENAME :

andyul00.

NO. OF POINTS	:	
25600		
WINDOW LENGTH (IN # OF SAMPLES)	:	
200		

WINDOW INCREMENT(IN # OF SAMPLES) :
 100
 SAMPLING FREQUENCY (IN HZ) :
 10000
 ORDER OF THE PREDICTOR :
 12
 WINDOW?
 RECT(0),HANN(1),HAMM(2),COSINE(3) :
 1
 AUTOCORR.(1) OR COVARIANCE(2) :
 1

TOTAL NO. OF FRAMES BEING ANALYSED : 250

AVERAGE MAGNITUDE OVER SILENCE INTERVAL: 10.8320

WISH TO COMPUTE PITCH CONTOUR (Y=1,N=0) :
 1

----- AUTOCORRELATION METHOD OF LP -----
 WINDOW LENGTH(# OF SAMPLES) : 200
 PREDICTOR ORDER : 12
 WINDOW INCREMENT (# OF SAMPLES) : 100
 NO. OF WINDOWS USED : 250

INPUT DATA FILENAME : andyul00.
 LP COEFFICIENT OUTPUT FILENAME : andyul00.LP
 SQUARED PRED ERROR OUTPUT FILENAME : andyul00.RS
 PITCH CONTOUR FILENAME : andyul00.PC

\$

\$ r wyesig

NOTE: YOU HAVE TO GIVE THE CORRECT VALUES FOR NUMBER
OF FRAMES AND NUMBER OF SAMPLES PER FRAME

GIVE THE TOTAL NUMBER OF FRAMES IN THE FILE:

250

GIVE NUMBER OF NONOVERLAPPING SAMPLES PER FRAME:

100

GIVE NUMBER OF SAMPLES PER FRAME:

200

OUTPUT FILE

enl100.

TO CHOOSE A MODEL FOR GLOTTAL PULSE,

TYPE [1] IF SINGLE IMPULSE EXCITATION

TYPE [2] IF DOUBLE IMPULSE EXCITATION

TYPE [3] IF TRIPLE IMPULSE EXCITATION

TYPE [4] FOR HILBERT TRANSFORM EXCTN

TYPE [5] FOR FANTS EXCITATION MODEL

TYPE [6] FOR LF EXCITATION MODEL

TYPE [7] FOR WHISPERED EXCITATION

TYPE [8] FOR MODEL 2 SINGLE PULSE EXCITATION

TYPE [9] FOR MODIFIED HILBERT TRANSFORM EXCTN

TYPE [10] FOR MODEL 2 LF EXCITATION

TYPE [11] FOR MODEL 3 LF EXCITATION

TYPE [12] FOR MODEL 3 SINGLE PULSE EXCITATION

TYPE [13] FOR MODEL 4 SINGLE PULSE EXCITATION

TYPE [14] FOR LF IMPULSE EXCITATION

TYPE [15] FOR LF IMPULSE EXCITATION 2

TYPE [16] FOR LF IMPULSE EXCITATION 3

WHAT IS YOUR CHOICE OF MODEL [1 OR 2 OR 3 OR ... 16]:

1

FILE NAME FOR PITCH PERIOD VALUES

andyul00.pc

FILE NAME FOR SQUARED GAIN VALUES

andyul00.rs

THE NUMBER OF FRAMES ARE: 448

THE TOTAL NUMBER SAMPLES WILL BE: 25005

THE TOTAL NUMBER OF EXCITATION SAMPLES ARE: 25005

1.000000

FORTRAN STOP

\$

```
$ r wuysy
LP COEFFFT FILENAME:
andyul00.lp
EXCITATION INPUT FILENAME :
enl100.
SYNTHESIZED OUTPUT FILENAME :
snw100.
PREDICTOR ORDER :
12
NUMBER OF SAMPLES TO BE GENERATED PER FRAME :
100
NO. OF FRAMES TO BE SYNTHESIZED:
250

    SYNTHETIC SPEECH GENERATED FOR FRAME:  251
IQR      25000
SYNTHESIS COMPLETED
$
```

APPENDIX D

HARDWARE

HARDWARE

The hardware required for our testing is shown in Figure 35. The natural speech is sensed with a microphone and then lowpass filtered before being sent to the A/D to be digitized. The digitized data is stored on a Vax 11/750 computer. After the synthesized signal has been created (in digitized

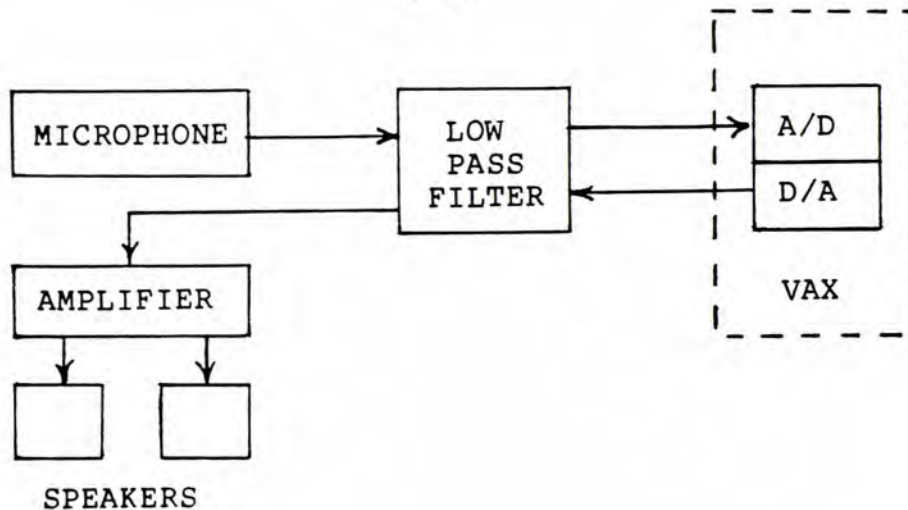


Figure 35. Hardware for speech tests.

form) it is sent to the D/A and lowpass filtered. This signal is then amplified and played through speakers or headphones.

Laboratory Peripheral Accelerator

The laboratory peripheral accelerator (LPA11-K) is an intelligent direct memory access (DMA) controller for laboratory data acquisition devices. It is used for applications requiring data acquisition at high rates. It contains 4 D/A's and 16 A/D's which interface between the computer memory and the audio equipment.

Lowpass Filter

The lowpass filters in this system are linear phase lowpass filters with cutoff frequencies of 5 kHz. Linear phase means that the time delay of all frequencies up to 5 kHz is the same or constant. The schematic is shown in figure 36. The phase is usually not of importance in filter design. However, if the time delay is not the same for all the components of the signal, then the time domain signal will be 'smeared', even though the frequency characteristics are unchanged. For example, a step function might be

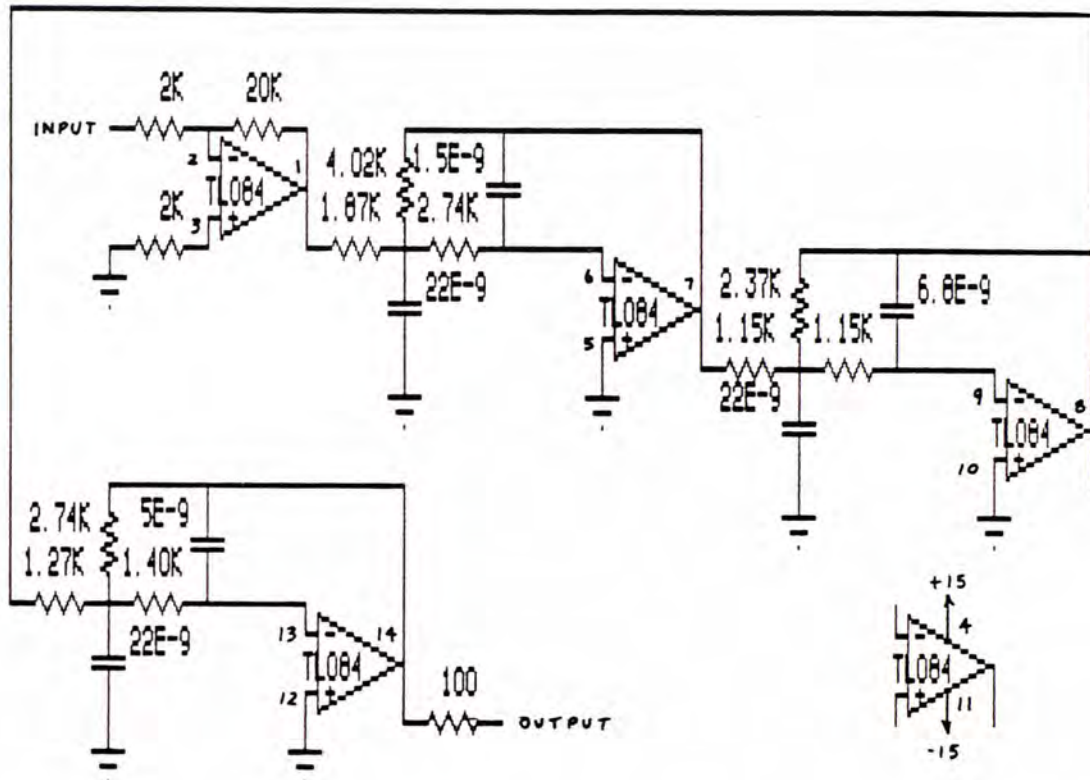


Figure 36. Schematic for linear phase lowpass filters.

smearred into a ramp. This smearing distorts short time speech events such as stops and is unacceptable for synthesizing quality speech.

Audio Equipment

The amplifier in Figure 1 is an Onkyo Tx-37 tuner amplifier. The microphone is an audiotechnica ATM31 unidirectional microphone. The speakers are speakerlab DAS3N digital audio speakers. All of these

components have a very good frequency response up to at least 5 KHz (the cutoff frequency for the lowpass filters).

APPENDIX E
EXCITATION PROGRAM

```

C          SIGNAL PROCESSING RESEARCH CENTER
C          DEPARTMENT OF ELECTRICAL ENGINEERING
C          &
C          COMMUNICATION SCIENCES
C          UNIVERSITY OF CENTRAL FLORIDA
C          ORLANDO, FL 32816
C          (305) 281-5786
C          DIRECTOR: Y. A. ALSAKA, PH. D., ASST. PROFESSOR
C*****
C 10      PROGRAM OR SUBROUTINE NAME: WUYESIG
C          AUTHOR(S)   : B. Yegnanarayana
C          DATE        : 5-9-83
C          MODIFIER(S) : Ke Wu, Andy Pearson
C          LAST MOD    : 12-11-86
C          VERSION     :
C          PURPOSE     : PROGRAM TO COMPUTE LPC EXCITATION SIGNAL
C
C          ABSTRACT    : COMPUTE ONE OF 16 EXCITATION SIGNALS AND PLACE
C                      IN AN ILS FILE.
C
C 20      SUBROUTINES AND FUNCTIONS REQUIRED (FOR LINKAGE): PACKB,CNVSD
C
C          REMARKS AND CONSTRAINTS:
C
C          DESCRIPTION OF PARAMETERS:
C              INPUT   : ANALYSIS PARAMETERS
C              OUTPUT  : ILS EXCITATION FILE WITH STANDARD HEADER
C
C          FORMAT OF FILES:
C              INPUT   :
C 3
C              OUTPUT  : ILS
C
C          ERRORS      :
C          SAMPLE CALL OR USAGE :R WUYESIG
C*****
C          DIMENSION P(500),G(500),GG(1000),Y(400),X(400),HT(500),XG(50000)
C          INTEGER*2 INFILE(12),IOUTFL(12)
C          INTEGER IIT1(500),IIT2(500),IT11(500)
C          INTEGER ITR2(500),IP(500),IQ(500)
C          CHARACTER*32 OFL
C          PI=3.141592653589793
C 1          TYPE*, 'NOTE: YOU HAVE TO GIVE THE CORRECT VALUES FOR NUMBER '
C          TYPE*, '      OF FRAMES AND NUMBER OF SAMPLES PER FRAME '
C          TYPE*, 'GIVE THE TOTAL NUMBER OF FRAMES IN THE FILE: '
C          ACCEPT*, NFR
C          TYPE*, 'GIVE NUMBER OF NONOVERLAPPING SAMPLES PER FRAME: '
C          ACCEPT*, NS
C          TYPE*, 'GIVE NUMBER OF SAMPLES PER FRAME: '
C          ACCEPT*, NSF
C          TYPE*, 'GIVE PITCH PERIOD MODIFICATION FACTOR [0.5 TO 2.0]: '
C          ACCEPT*, AL
C          AL=1.0
C          TYPE*, 'OUTPUT FILE '
C          READ(*,39)OFL
C 34          FORMAT(A)

```

```

TYPE*, 'TO CHOOSE A MODEL FOR GLOTTAL PULSE, '
TYPE*, ' TYPE [1] IF SINGLE IMPULSE EXCITATION'
TYPE*, ' TYPE [2] IF DOUBLE IMPULSE EXCITATION'
TYPE*, ' TYPE [3] IF TRIPLE IMPULSE EXCITATION'
TYPE*, ' TYPE [4] FOR HILBERT TRANSFORM EXCTN '
TYPE*, ' TYPE [5] FOR FANTS EXCITATION MODEL '
TYPE*, ' TYPE [6] FOR LF EXCITATION MODEL '
TYPE*, ' TYPE [7] FOR WHISPERED EXCITATION '
TYPE*, ' TYPE [8] FOR MODEL 2 SINGLE PULSE EXCITATION'
TYPE*, ' TYPE [9] FOR MODIFIED HILBERT TRANSFORM EXCTN'
TYPE*, ' TYPE [10] FOR MODEL 2 LF EXCITATION'
TYPE*, ' TYPE [11] FOR MODEL 3 LF EXCITATION'
TYPE*, ' TYPE [12] FOR MODEL 3 SINGLE PULSE EXCITATION'
TYPE*, ' TYPE [13] FOR MODEL 4 SINGLE PULSE EXCITATION'
TYPE*, ' TYPE [14] FOR LF IMPULSE EXCITATION 2'
TYPE*, ' TYPE [15] FOR LF IMPULSE EXCITATION '
TYPE*, ' '
TYPE*, 'WHAT IS YOUR CHOICE OF MODEL
$ [1 OR 2 OR 3 OR ... 16]: '
ACCEPT*, IMOD
IF ((IMOD .GT. 18) .OR. (IMOD .LT. 1)) GO TO 111
IF ((IMOD .EQ. 1) .OR. (IMOD .EQ. 4)) GO TO 112
IF ((IMOD .EQ. 6) .OR. (IMOD .EQ. 10)) GOTO 1112
IF ((IMOD .EQ. 11) .OR. (IMOD .GT. 14)) GO TO 1112
IF ((IMOD .EQ. 12) .OR. (IMOD .EQ. 9)) GO TO 112
IF ((IMOD .EQ. 8) .OR. (IMOD .EQ. 13)) GO TO 112
NOY=1
IF (NOY.EQ.0) GOTO 110
ITT1=40
ITT2=20
IF ((ITT2 .LT. 5) .OR. (ITT2 .GT. 30)) GO TO 111
IF ((ITT1 .LT. 30) .OR. (ITT1 .GT. 60)) GO TO 111
GO TO 112
111 TYPE*, ' YOU MAY HAVE TYPED A WRONG NUMBER! TRY AGAIN: '
GO TO 1
100 FORMAT(12A2)
110 TYPE*, 'FILE NAME FOR ARRAYS OF T1 AND T2 VALUES : '
READ 100, INFILE
OPEN (4, FILE=INFILE, STATUS='OLD', FORM='UNFORMATTED',
$ RECORDTYPE='FIXED', RECL=1)
DO 115 J=1, NFR
READ (4) IT1(J)
115 READ (4) IT2(J)
GOTO 112
1112 INTP=40
INTE=50
INTA=3
IF (IMOD .EQ. 15) INTP=46
IF (IMOD .EQ. 15) INTE=58
FI4= 6
IF((IMOD.EQ.6) .OR. (IMOD.GT.14))GO TO 3334
TYPE*, '# OF ZEROS BETWEEN PULSES ?'
ACCEPT*, NZER
NZER=NZER+1
IF(IMOD.EQ.10) GO TO 3334
TYPE*, '# TO INCREMENT'
ACCEPT*, NINC
3334 RA=INTA/FLOAT(100-INTE)
IF(RA.LE.0.5.AND.RA.GE.0.1) AK=2-2.34*RA*RA+1.34*RA*RA*RA*RA
IF(RA.GT.0.5) AK=2.16-1.32*RA+0.64*(RA-0.5)*(RA-0.5)

```

```

IF(RA.LT.0.1) AK=2.0
UE=0.5*INTA*AK/INTP
ST=SIN(INTE*PI/INTP)
CT=COS(INTE*PI/INTP)
112 TYPE*, 'FILE NAME FOR PITCH PERIOD VALUES'
READ 100, INFILE
IF(IMOD.NE.7) GO TO 1113
TYPE*, 'GIVE VALUE OF SNR FOR MODEL 6(WHISPER) [0. TO 1.0]: '
ACCEPT*, SNR
1113 OPEN (1, FILE=INFILE, STATUS='OLD', FORM='UNFORMATTED',
$ RECORDTYPE='FIXED', RECL=1)
DO 116 J=1, NFR
116 READ (1) P(J)
TYPE*, 'FILE NAME FOR SQUARED GAIN VALUES'
READ 100, INFILE
OPEN (2, FILE=INFILE, STATUS='OLD', FORM='UNFORMATTED',
$ RECORDTYPE='FIXED', RECL=1)
DO 117 J=1, NFR
117 READ (2) G(J)
C INITIALIZATION OF HT ARRAY
C
DO 55 JJ=1, 300, 2
HT(JJ) = 2. / (PI*JJ)
HT(JJ+1) = 0.
55 CONTINUE
C
C GENERATE NEW PITCH AND GAIN CONTOUR BY COMPUTING THE PITCH AND
C GAIN AT THE END OF EACH PITCH PERIOD
C
DO 10 I=1, NFR
C TYPE *, I, IP(I), AL, P(I)
IP(I) = AL*P(I)
10 CONTINUE
G(1) = G(1)/NSF
NTT = NS*NFR
J = 1
IP(1) = 0
G(1) = 0
GG(1) = G(1)
IQ(1) = IP(1)
IF (NOY.EQ.1) GOTO 13
IIT1(1)=IT11(1)
IIT2(1)=IT22(1)
13 ISUM = NS
113 J = J+1
IL = ISUM/NS
IIQ = IQ(J-1)
GGG = GG(J-1)
300 FORMAT(2X, 'J=', I3, ' IL=', I3, ' IQ(J-1)=', I3, ' ISUM=', I5)
C
C NOTE: NS IS THE NUMBER OF NEW SAMPLES IN EACH ANALYSIS FRAME
C
IF (IP(IL+1) .GT. 0) GO TO 11
IQ(J) = IP(IL+1)
GG(J) = G(IL+1)
IF (NOY.EQ.1) GOTO 14
IIT1(J)=IT11(IL+1)
IIT2(J)=IT22(IL+1)
14 ISUM = ISUM + NS
IF (ISUM .GT. NTT) GO TO 999

```



```

GO TO 113
11 IF (IP(IL) .GT. 0) GO TO 12
   IQ(J) = IP(IL+1)
   GG(J) = G(IL+1)
   IF (NOY.EQ.1) GOTO 15
   IIT1(J)=IT11(IL+1)
   IIT2(J)=IT22(IL+1)
15 ISUM = ISUM + IQ(J)
   IF (ISUM .GT. NTT) GO TO 999
GO TO 113
12 INC = ISUM - IL*NS
   FFLONS = FLOAT(NS)
   IQ(J) = IP(IL) + ((IP(IL+1) - IP(IL))*INC)/FFLONS
   GG(J) = G(IL) + ((G(IL+1) - G(IL))*INC)/FFLONS
   IF (NOY.EQ.1) GOTO 16
   IIT1(J)=IT11(IL)+((IT11(IL+1)-IT11(IL))*INC)/FFLONS
   IIT2(J)=IT22(IL)+((IT22(IL+1)-IT22(IL))*INC)/FFLONS
16 ISUM = ISUM + IQ(J)
   IF (ISUM .GT. NTT) GO TO 999
GO TO 113
999 TYPE 200, J, ISUM
200 FORMAT(10X, 'THE NUMBER OF FRAMES ARE: ', I6// ' THE TOTAL NUMBER OF
   * SAMPLES WILL BE: ', I10)
C
C FROM THIS POINT TILL THE END IT IS GENERATION OF EXCITATION SIGNAL
C
   NNPTS = ISUM
   NNFR = J
   ISIG = 0
   ISUM = 0
   IX = 551
C
C NOTE: AK DETERMINES THE COMPONENT OF RANDOM NOISE TO BE ADDED
C MINIMUM VALUE OF AK=0.01
C
   AK = 0.01
   DO 40 II=1, NNFR
   IF (NOY.EQ.1) GOTO 23
   IIT1=IIT1(II)
   IIT2=IIT2(II)
23 TT1=IIT1/100.
   TT2=IIT2/100.
   KK = IQ(II)
   IF (KK .LE. 0) GO TO 777
C-----
7465 B = SQRT(GG(II))
7466 BK = AK*B
C----
   T1 = TT1*KK
   T2 = TT2*KK
   IT1 = T1
   IT2 = T2
C
C GENERATE RANDOM NOISE SAMPLES
C
   TYPE *, 'KK', KK
   DO 45 JJ=1, KK
   IY = IX*35
   IY = MOD(IY, 929)
   YFL = FLOAT(IY)

```

```

YFL = YFL/929
Y(JJ) = 2*(YFL-0.5)
IX = IY
45 CONTINUE
SUM = 0.
DO 50 JJ = 1, KK
Y(JJ) = BK*Y(JJ)
SUM = SUM + Y(JJ)*Y(JJ)
50 CONTINUE
GO TO (101, 102, 103, 104, 105, 106, 107, 108, 104, 106, 106, 108,
1108, 106, 106, 106, 106), IMOD
C
C EXCITATION MODEL 1
C
101 CONTINUE
IF(KK.LE.1)A=1.0
IF(KK.LE.1) GO TO 1011
A = 1/FLOAT(KK-1)
C-----
1011 CCC = ((GG(II)*KK-SUM)/((KK-1)*A*A + 1))
C IF(CCC.GE.0) GO TO 2039
C CC=0.
C GO TO 2040
2039 CC = SGRT(CCC)
2040 DO 60 JJ=1, KK
C----
X(JJ) = -A*CC + Y(JJ)
60 CONTINUE
X(KK) = CC - A*CC
GO TO 888
C
C EXCITATION MODEL 8, 12 AND 13
C
108 CONTINUE
IF(KK.LE.1)A=1.0
IF(KK.LE.1) GO TO 1811
A = 1/FLOAT(KK-1)
1811 CC = SGRT((GG(II)*KK-SUM)/((KK-1)*A*A + 1))
DO 1860 JJ=1, KK
X(JJ) = -A*CC + Y(JJ)
1860 CONTINUE
X(KK) = CC - A*CC
IF ((IMOD.EQ.12).OR.(IMOD.EQ.13)) GO TO 1212
X(KK-2)=X(KK)/2
1212 X(KK-1)=X(KK)
X(KK)=X(KK)/2
IF(IMOD.EQ.13) GO TO 1213
GO TO 888
1213 X(KK)=X(KK-1)
GO TO 888
C
C EXCITATION MODEL 2
C
102 CONTINUE
IF(KK.LE.2) A=1.0
IF(KK.LE.2) GO TO 1738
A = 1/FLOAT(KK-2)
1738 CC = SGRT((GG(II)*KK-SUM)/((KK-2)*A*A + 1 + (T2/T1)*(T2/T1)))
DO 70 JJ=1, KK
X(JJ) = -A*CC + Y(JJ)

```

```

70      CONTINUE
        X(KK) = CC - A*CC
        X(KK-IT1-IT2) = CC*((T2/T1)-A)
        GO TO 888

C
C      EXCITATION MODEL 3
C
103     T = T2/T1
        TT = (T1+T2)/T1
        CC = SGRT((GG(II)*KK-SUM)/(1+T*T+TT*TT))
        DO 80 JJ=1, KK
          X(JJ) = Y(JJ)
80      CONTINUE
        X(KK) = CC + X(KK)
        X(KK-IT2) = -CC*TT + X(KK-IT2)
        X(KK-IT1-IT2) = CC*T + X(KK-IT1-IT2)
        GO TO 888

C
C      EXCITATION MODEL 4 AND 9
C
104     X(1) = 0.
        X(2) = 0.
        K2 = KK/2 + 1
        K22 = 2*K2
        K21 = K2 + 1
        X(K2) = 0.
        PSUM = 0.01
        DO 81 JJ=K21, KK
          X(JJ) = HT(JJ-K21+1)
          X(K22-JJ) = -X(JJ)
          PSUM=PSUM + 2*X(JJ)*X(JJ)
81      CONTINUE
        CC = GG(II)*KK/PSUM
1081    CC = SGRT(CC)
        DO 82 JJ=1, KK
          X(JJ) = X(JJ)*CC
82      CONTINUE
        IF(IMOD.EQ.4) GO TO 888
        DO 1047 JJ=1, KK
1047    IF(X(JJ).EQ.0) X(JJ)=(X(JJ-1)+X(JJ+1))/2
        GO TO 888

C
C      EXCITATION MODEL 5
C
105     FWG = PI/T1
        IFT1 = KK-IT1-IT2
        IFT11 = IFT1+1
        IFT2 = KK-IT2
        IFT21 = IFT2+1
        IFT3 = KK
        FK = COS(FWG*T2)
        FK = 1./(1-FK)
        DO 830 JJ=1, IFT1
830     X(JJ) = 0.
        SUM1 = 0.01
        DO 831 JJ=IFT11, IFT2
          X(JJ) = 0.5*FWG*SIN(FWG*(JJ-IFT1))
831     SUM1 = SUM1 + X(JJ)
        SUM2 = 0.
        DO 832 JJ=IFT21, IFT3

```

```

      X(JJ) = -FK*FWG*SIN(FWG*(JJ-IFT2))
832  SUM2 = SUM2 + X(JJ)
      C
      C   FOR ZERO MEAN EXCITATION
      C
      FSUM = -SUM2/SUM1
      DO 833 JJ=IFT1, IFT2
833  X(JJ) = FSUM*X(JJ)
      SUM1 = 0.01
      DO 834 JJ=1, IFT3
      SUM1 = SUM1 + X(JJ)*X(JJ)
834  CONTINUE
      CC = SQRT((GG(II)*KK - SUM)/SUM1)
      DO 835 JJ=1, KK
835  X(JJ) = CC*X(JJ) + Y(JJ)
      GO TO 888

      C
      C   EXCITATION MODEL 6, 10, 11, 14, 15, AND 16
      C
106  ITC=KK
      ITP=NINT(INTP*KK/100.)
      ITE=NINT(INTE*KK/100.)
      TA=INTA*KK/100.
      IF(ITP.EQ.0) ITP=1
      WG=PI/ITP
      WU=WG*(WG*UE-CT/ST/ITP)

      C
      C   WEGSTEIN'S METHOD TO SOLVE FOR A
      C
      AO=0.5*WG
      EAT=EXP(-AO*ITE)
      FO=-WG*EAT/ST-(AO*AO*UE+WU)*ITP
      A1=FO
830  EAT=EXP(-A1*ITE)
      F1=-WG*EAT/ST-(A1*A1*UE+WU)*ITP
      AQ=A1-F1
      IF(AQ.EQ.0) GO TO 640
9876 A=A1+(A1-AO)/((AO-FO)/(A1-F1)-1)
9875 IF(ABS(A1-A).LE.1.0E-6) GOTO 640
      FO=F1
      AO=A1
      A1=A
      GOTO 630

      C
      C   NEWTON-PAPHSON METHOD TO SOLVE FOR EPS
      C
640  EPSX=1.0/TA
      EPS=EPSX
      CE=ITC-ITE
850  ECE=EXP(-EPS*CE)
      EPS=EPS-(EPS*TA+ECE-1)/(TA-ECE*CE)
      IF(ABS(EPSX-EPS).LE.1.0E-6) GOTO 655
      EPSX=EPS
      GOTO 650

      C
      C   COMPUTE EXCITATION WAVEFORM
      C
655  EEO=500.
      EO=-EEO*EAT/ST

```

```

SUM1=0. 01
DO 610 JJ=1, ITE+1
610 X(JJ)=EO*EXP(A*(JJ-1))*SIN(WG*(JJ-1))
SUM1=SUM1+X(JJ)
IFT1=ITE+2
IF(EPS. EQ. 0) EPS=. 01
EET=-EEO/(EPS*TA)
SUM2=0. 0
DO 620 JJ=IFT1, ITC
620 X(JJ)=EET*(EXP(-EPS*(JJ-ITE-1))-ECE)
SUM2=SUM2+X(JJ)
C
C FOR ZERO MEAN EXCITATION
C
FSUM = -SUM2/SUM1
DO 633 JJ=1, ITE+1
633 X(JJ) = FSUM*X(JJ)
SUM1 = 0. 01
DO 634 JJ=1, KK
SUM1 = SUM1 + X(JJ)*X(JJ)
634 CONTINUE
CC = SQRT((GG(II)*KK - SUM)/SUM1)
DO 635 JJ=1, KK
635 X(JJ) = CC*X(JJ) + Y(JJ)
IF(IMOD. EQ. 6) GO TO 888
IF(IMOD. GT. 14) GO TO 1414
IF(IMOD. EQ. 11) GO TO 1173
KK3=KK/NZER
DO 1067 JJ=1, KK3
JH1=(JJ-1)*NZER+2
JH2=JJ*NZER
DO 1068 JH=JH1, JH2
1068 X(JH)=0.

1067 CONTINUE
GO TO 888
1173 CONTINUE
C KK-POINTS IN PITCH PERIOD, NZER-# OF ZEROS BETWEEN FIRST 2 IMPULSES
C NINC-# MORE ZEROS BETWEEN 1 AND 2 IMPULSE THAN BETWEEN 2 AND 3
KKQ=KK-1
NZERM=NZER-1
1144 KKQ1=0
CCCC INLOOP
1146 CONTINUE
X(KKQ)=0.
KKQ=KKQ-1
IF(KKQ. LE. 0) GO TO 888
KKQ1=KKQ1+1
IF(KKQ1. NE. NZERM) GO TO 1146
NZERM=NZERM+NINC
KKQ=KKQ-1
GO TO 1149
1414 I1=KK*. 15
I2=KK*. 34
I3=KK*. 44
I4=KK*F I4
1438 I5=KK*. 67
DO 1472 I7=1, KK
1472 X(I7)=X(I7)*5
DO 1400 I7=1, I1-1

```

```

1400 X(I7)=0.
      DO 1401 I7=I1+1, I2-1
1401 X(I7)=0.
      DO 1402 I7=I2+1, I3-1
1402 X(I7)=0.
      DO 1403 I7=I3+1, I4-1
1403 X(I7)=0.
      DO 1404 I7=I4+1, I5-1
1404 X(I7)=0.
      DO 1406 I7=I5+1, KK
1406 X(I7)=0.
      GO TO 888
C
C      EXCITATION MODEL 7
C
107  FWG = PI/T1
      IFT1 = KK-IT1-IT2-IT2
      IFT11 = IFT1+1
      IFT2= KK-IT2-IT2
      IFT21 = IFT2+1
      IFT3= KK-IT2
      FK = COS(FWG*T2)
      FK = 1./(1-FK)
      DO 840 JJ=1, IFT1
140  X(JJ) = 0.
CCC
      SUM1 = 0.01
CCC
      DO 841 JJ=IFT11, IFT2
141  X(JJ) = 0.5*FWG*SIN(FWG*(JJ-IFT1))
      SUM1 = SUM1 + X(JJ)
      SUM2 = 0.
      DO 842 JJ=IFT2, IFT3
142  X(JJ) = -FK*FWG*SIN(FWG*(JJ-IFT2))
      X(KK-JJ+IFT2) = X(JJ)
      SUM2 = SUM2 + X(JJ)
      SUM2 = 2*SUM2 - X(IFT3)
C
C      FOR ZERO MEAN EXCITATION
C
      FSUM = -SUM2/SUM1
      DO 843 JJ=IFT11, IFT2
143  X(JJ) = FSUM*X(JJ)
CCC  HAD PROBLEM WITH DIVIDING BY ZERO
      SUM1 = 0.01
CCC
      DO 844 JJ=1, KK
144  SUM1 = SUM1 + X(JJ)*X(JJ)
      CONTINUE
      CC = SQRT((GG(II)*KK)/(2.*SUM1))
      CC1 = SQRT(SUM1/SUM)
      SNR1 = SQRT(1.-SNR*SNR)
      DO 845 JJ=1, KK
145  X(JJ) = 1.414*CC*(SNR*X(JJ) + SNR1*CC1*Y(JJ))
      GO TO 888
C
C      GENERATE RANDOM NOISE FOR UNVOICED EXCITATION
C
777  SUM = 0.
C      TYPE *, 'NS', NS

```

```

      DO 85 JJ=1, NS
      IY = IX*35
      IY = MOD(IY, 929)
      YFL = FLOAT(IY)
      YFL = YFL/929
      Y(JJ) = 2*(YFL-0.5)
      IX = IY
      SUM = SUM + Y(JJ)*Y(JJ)
85    CONTINUE
      SUM = SUM/NS

C-----
      CCC = ((GG(II))/SUM)
9087  CC=SQRT(CCC)
9088  DO 90 JJ=1, NS
      X(JJ) = CC*Y(JJ)
90    CONTINUE
C
C      WRITE THE EXCITATION SIGNAL FOR THIS PERIOD ON THE OUTPUT FILE
C
888  IF (KK LE. 0) KK=NS
      DO 889 JJ=1, KK
889  XG(ISUM+JJ)=X(JJ)
      ISUM = ISUM + KK
400  FORMAT (2X, 'THE TOTAL NUMBER OF EXCITATION SAMPLES ARE: ', I10)
40    CONTINUE
      TYPE 400, ISUM
      CALL CNVSUD(XG, ISUM, OFL)
      STOP
      END

```

REFERENCES

- Atal, B. S., and Hanauer, S.L. "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave." *Journal of the Acoustical Society of America* 50 (1971): 637-655.
- Borden, G. J., and Harris, K. S. Speech Science Primer. Baltimore: William and Wilkins Co., 1980.
- Childers, D. G.; Hicks, D. M.; Moore, G. P.; and Alsaka, Y. A. "A Model For Vocal Fold Vibratory Motion, Contact Area, and the Electroglossogram." *Journal of the Acoustical Society of America* 80 (1986): 1309-1320.
- Childers, D. G. and Wu, Ke. "Some Factors Responsible For Quality, Intelligibility, and Naturalness of Synthetic Speech." Paper presented at the University of Florida, Gainesville, Fl., 1986.
- Childers, D. G.; Yea, J. J.; and Krishnamurthy, A. "Spectral Analysis: AR, MA, ARMA." Paper presented at the University of Florida, Gainesville, Fl., 1986.
- Fant, G. "Vocal Source Analysis - A Progress Report." *STL-QPSR* 3-4 (1979): 31-53.
- Fant, G.; Liljencrants, J.; and Lin, Q. G. "A Four Parameter Model of Glottal Flow." Paper presented at the Vocal Fold Physiology Symposium, New Haven, Conn., 1985.
- Flanagan, J. L.; Coker, C. H.; Rabiner, L. R.; Schafer, R. W.; and Umeda, N. "Synthetic Voices For Computers." *IEEE Spectrum* 7 (1970): 22-45.
- Jain, V. K. "Multi-Pulse Spectral Analysis Of Speech." *IEEE Spectrum* (1983): 80-83.
- Kaplan, Gadi and Lerner, E. J. "Realism in Synthetic Speech." *IEEE Spectrum* (1985): 32-37.
- Klatt, D. H. "Software For a Cascade/Parallel Formant Synthesizer." *Journal of the Acoustical Society of America* 67 (1980): 971-995.

- Larar, J. N., and Childers, D. G. "Spectral Analysis of Speech and EGG Using Pitch Synchronous Windows Over One Period." Paper presented at the University of Florida, Gainesville, Fl., 1986.
- Makhoul, John. "Linear Prediction: A Tutorial Review." Proceedings of the IEEE 63 (1975): 561-580.
- Naik, J. M. "Synthesis And Evaluation Of Natural Sounding Speech Using The Linear Predictive Analysis-Synthesis Scheme." Ph.D. dissertation, University of Florida, 1984.
- Oppenheim, A. V. Applications Of Digital Signal Processing. Englewood Cliffs, New Jersey: Prentice Hall, 1978.
- Pearson, W. A. "All-Pole Filter Model For LPC Speech Production." Paper presented at the University of Central Florida, Orlando, Fl., 1986.
- Wu, Ke "A Flexible Speech Analysis-Synthesis System For Voice Conversion." Masters thesis, University of Florida, 1985.