NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

ISBN 0-315-53225-4

Canada

IMPROVEMENTS TO A PITCH-SYNCHRONOUS LINEAR PREDICTIVE CODING (LPC) VOCODER

by

Raymond C. Carr, B.A.Sc.

A thesis submitted to the
School of Graduate Studies and Research
in partial fulfillment of the requirements
for the degree of

Master of Applied Science

Ottawa-Carleton Institute for Electrical Engineering
Department of Electrical Engineering
Faculty of Engineering
August, 1989

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

CONTENTS

ABSTRACT

This thesis deals with careful non-real time LPC analysis. First, a baseline system is described. It uses a pitch-synchronous covariance method analysis, with the pitch-synchrony provided by a laryngograph signal. The reliability of voicing decisions and fundamental frequency determination is increased, and work to find a better voiced excitation waveform is then presented. Buzziness in the synthesis of voiced-fricatives is reduced by adding white noise to the excitation, and regions where this should be done are automatically determined using three parameters: the total power, the ratio of high frequency power to total power, and the bandwidth of the first formant. Properties of covariance method LPC analysis are explored: setting a lower limit on the value of B_1 is found to increase the quality of the synthesis, and the need for pitch-synchrony is discussed.

ACKNOWLEDGMENTS

I thank my supervisor, Dr. Melvyn J. Hunt, for his valuable guidance, and for the time he spent correcting the thesis.

My thanks are also directed at the National research Council of Canada, in particular to Dr. S.R.M. Sinclair, head of the NAE Flight Research Laboratory, for the use of computer resources, and to Carl P. Swail and Claude Lefèbvre, for their help with the computer system.

I also thank Dariusz A. Zwierzynski, from the University of Ottawa Linguistics Department, for his feedback on the thesis, and for the time he spent preparing test material for the research described in this thesis. In addition, I thank Dr. Peter Galko, from the Electrical Engineering Department, for his comments on the thesis.

I also acknowledge financial support from the NSERC Operating Research Grant No. C48 6681 issued to Dr. Melvyn Hunt, and from the Department of National Defence through DCIEM for partial funding of the computing facilities at the NAE Flight Research Laboratory.

SUMMARY

Electronic speech analysis and synthesis systems based on the source-filter model for speech production can be traced back to 1939, when H. Dudley at Bell Laboratories developed the *voice coder*, or *vocoder*[1]. This device, more precisely referred to as a *channel vocoder*, used a bank of bandpass filters to obtain a representation of the amplitude spectrum of the speech signal. An excitation signal, either a periodic signal or white noise, was passed through a bank of bandpass filters and the outputs summed to produce synthetic speech. The main motivations at the time, and probably to this day, were fundamental interests in speech production and efficient voice communications.

The advent of digital computers, sampled-data theory and integrated circuits created new possibilities for speech synthesis. Motivations for speech output research from the 1960s to the present include low bit-rate digital speech communication systems and text-to-speech systems. More recent applications for speech synthesis (and speech recognition) include air traffic control equipment, aids for the handicapped, and study of the human auditory system and its disorders. Speech input/output is also particularly attractive in aeronautical applications because the workload on fighter and helicopter pilots has increased tremendously during the past decades, and speech input/output can relieve pilots of manual tasks. Such is the motivation for the research being conducted and the National Aeronautical Establishment (NAE) Speech Research Centre, a division of the National Research Council of Canada (NRC). More specifically, regarding

speech output, the aim of the Centre is to produced synthetic speech of high quality and high flexibility. High quality speech simply means that the speech is intelligible without effort. This is important in applications that involve human risk, as tests show that our reaction time to unnatural speech is longer than our reaction time to natural speech[2]. Flexibility means that timing and intonation can be changed; this is useful to add emotions to the speech, such as making it sound urgent, or giving it a more relaxed tone. Also, different voices can be used for different sources of information. To achieve these goals, the Centre developed a Linear Predictive Coding (LPC) based speech analysis and synthesis system[3]. The Centre has also successfuliy developed algorithms to manipulate voice characteristics.

LPC is a waveform coding technique that is commonly used for low-bit rate digital speech communications. In such applications, several factors limit the quality of the reproduced speech. For example, the analysis and synthesis must be carried out in real-time, thus limiting the complexity of the algorithms that can be used, and real-time communication systems must often operate with speech degraded by noise, reverberation, and distortions, which degrade the quality of the LPC analysis.

There are, however, other applications for which these limitations need not apply. LPC may be used to estimate formants (model the vocal tract), to generate stimuli for speech perception experiments, and to encode speech efficiently, as in the system being developed at the Centre. For these applications, the analysis need not function in real time, and the acoustic quality of the speech

to be analyzed (and reproduced) can be carefully controlled. As a result of this, synthetic speech of high quality can be obtained.

Although the speech quality of the LPC system developed at the Centre was high, it was suspected that it could be improved; this was the aim of the research described here. The author first verified various parameters and assumptions about the LPC-based system, such as the amount of speech preemphasis, quantization of filter coefficients, coding of filter coefficients, analysis order, smoothing of covariance matrices, and excitation function used. Having determined that the assumptions were valid and that the parameters were set correctly, careful listening was used to identify specific problems.

The first three chapters of this thesis present introductory information: Chapter 1 is on speech sounds in general, Chapter 2 presents speech output systems, with an emphasis on vocoder-based systems, and Chapter 3 introduces the LPC-based vocoder developed at the Centre.

The research that was carried out by the author is then presented in the remaining chapters. Chapter 4 presents problems caused by the original speech segmentation program and how they were resolved, whereas Chapter 5 deals with excitation functions for voiced speech. The synthesis of frames with mixed-excitation is the subject of Chapter 6, and aspects of pitch-asynchronous LPC are discussed in Chapter 7. This chapter does not deal with solving a specific problem to improve synthesis, but rather presents properties of LPC analysis that were discovered during this research, and that may have wider implications.

Evaluation of the performance of high quality LPC analysis and synthesis is not easy. There exists several methods to obtain objective measures of speech quality[4]. Some measure the accuracy in fitting spectra or locating formants; these methods require a standard of reference for what constitutes "correct" spectra or formant data for real speech. Other methods measure continuity of formant tracks. Such methods may give an indication of the reliability of the analysis, but may also be misleading, because analysis methods which use long overlapping window naturally impose continuity on the formant tracks. The normalized residual power can also be used, but this measure does not always correlate with subjective judgments of quality. Signal-to-noise ratios, often used to evaluate waveform coding schemes, are not useful here because LPC-encoded waveforms can show differences from the original speech that are not perceptible. Although we have not confirmed this belief experimentally, the intelligibility of the synthesized speech of the described system is almost certainly too high to be measured objectively. For these reasons, informal subjective testing was used: the author listened to various versions of a given synthetic speech file and chose the version that sounded the best. In case of a doubt, several listeners took part in a blind test to come to a conclusion.

CHAPTER 1

GENERAL INFORMATION ON SPEECH SOUNDS

1.1 Physiological aspects of speech production

Several organs and muscles are involved in the speech production process. The most important ones are the diaphragm, chest muscles, lungs, windpipe, vocal cords, throat, nasal passages, sinuses, mouth, and jaw. The diaphragm sets the whole operation in motion. It is a flat muscular membrane below the rib cage, and it separates the chest cavity from the abdomen. When we breathe, the diaphragm lowers, and our chest muscles move our ribs outwards, thus distending our lungs. This causes air to flow in. We then contract the diaphragm and our rib cage, causing some of the air in the lungs to flow out. The air passes through the larynx, into the throat, nasal passages, sinuses, mouth, and out into the open. We produce acoustic energy when we obstruct this air flow; this is usually done with our vocal cords, but other mechanisms are possible. Once acoustic energy is created, resonance from the chest, throat, mouth and nose determines the identity of the sound we hear.

The *vocal cords*, or perhaps more technically the *vocal folds*, play a major role in the production of speech sounds. Located at the top of the trachea in an organ called the *larynx*, the vocal cords are two muscular folds that lie opposite to each other. When relaxed, they are opened and do not obstruct the

air flow. When tensed, they come together, and are capable of a complete blocking of the air flow, such as when we hold our breath. The space between the vocal folds is called the *glottis*, and the "glottal air flow" refers to the air volume that flows past the cords as a function of time. To produce acoustic energy, the vocal cords undergo a periodic movement, and a single cycle of this movement, called a glottal cycle, can be described as follows: the cords are initially closed; air pressure from the lungs forces them to open, and as air starts to flow, the air pressure in the glottis drops (Bernouilli principle). The low pressure violently sucks the cords together, producing acoustic energy. The air pressure behind the closed cords then starts to build up again, and the process repeats itself. The vocal cords can in some ways be compared to the lips, which can block the air flow to various degrees, and can be made to open and close in a fast periodic way (an example of such a fast periodic motion is the *B'rrr* sound —a labial trill— that we make to indicate coldness). The *closed glottis* interval of a glottal cycle is when the folds are closed and completely block the air flow from the lungs. The open glottis interval is when the folds are opened. Figure 1 shows the air flow volume through the glottis as a function of time. Although acoustic energy can be generated at other times in the glottal cycle[5], as occurs in breathy or creaky phonation, it is normally concentrated at the instant the vocal cords close due to the large rate-of-change of the glottal air flow at this instant.

The rate at which the vocal cords open and close is called the *fundamental frequency*, and it correlates strongly with the perceived pitch of a speech sound. The fundamental frequency is typically 100 times per second for male

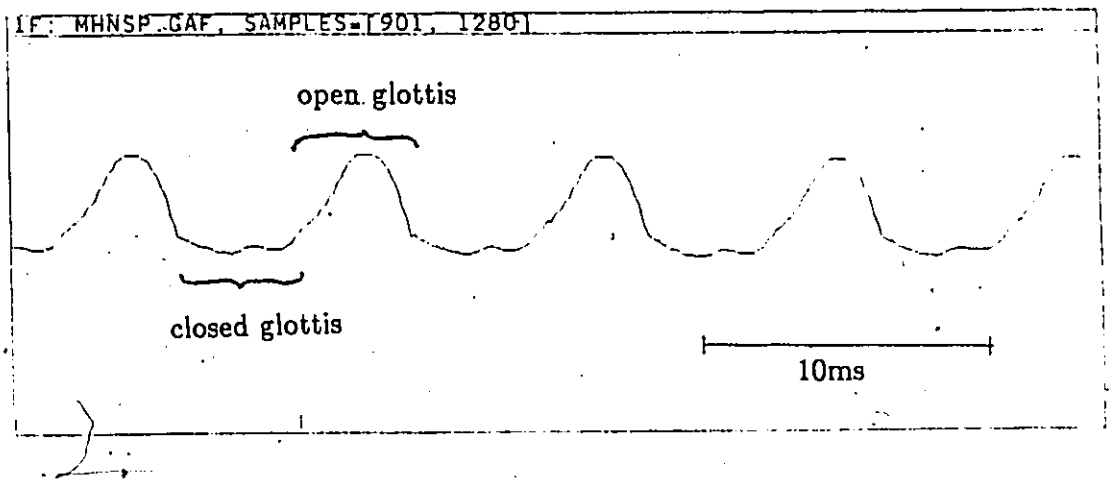

Figure 1: Glottal air flow, showing the closed glottis and open glottis intervals.

speech and 200 times per second for female speech. Determined by the air pressure in the trachea and by physiological characteristics of the vocal cords such as length, thickness, and tension, the fundamental frequency changes slowly relative to the movement of other speech organs.

During normal speech, the amplitude spectrum of the glottal air flow consists of a set of harmonics separated from each other by the fundamental frequency. The spectrum envelope falls of at an approximate rate of -12 dB/octave (see Figure 2).

To some extent, we can change the spectrum of the sound generated by the vocal cords. This can be done, for example, by modifying the tenseness of the cords, or by varying the amount of air we expel.

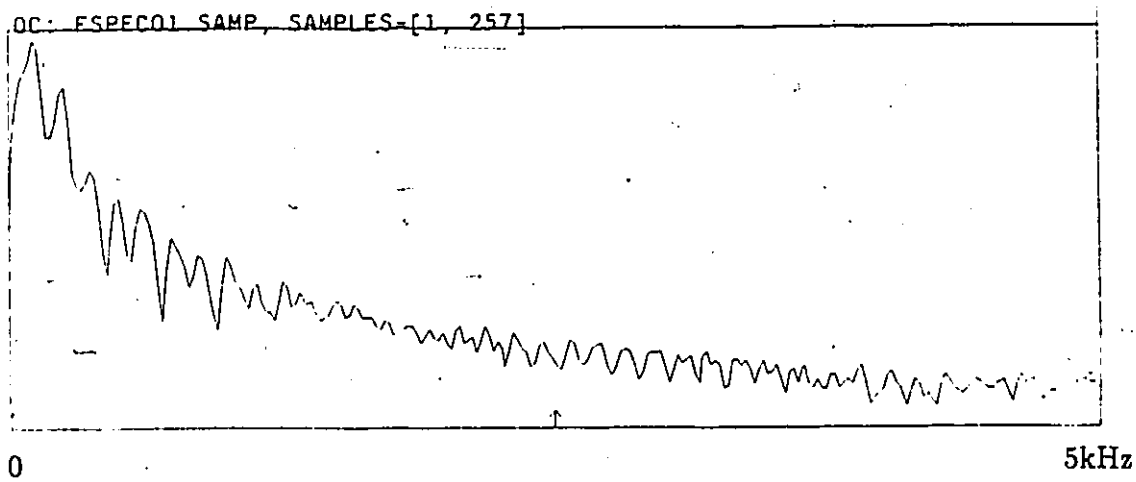

Figure 2: Log amplitude spectrum of glottal air flow

1.2 Classification of speech sounds

Speech sounds can be broadly classified according to how acoustic energy is produced. When the only source of acoustic energy is the vocal cords, the sound is classified as *voiced*. Examples of this are the vowels in *fat*, *food*, and *feet*, and many consonants such as /l/, /m/, and /n/. (The / / symbol is used to indicate phonemes, which will be described in the next paragraph.) When the vocal cords are not vibrating, but the articulators — teeth, tongue, lips — or half-closed cords are used to create an audible turbulence in the air flow (frication), the sound is said to be *voiceless*. Examples of voiceless sounds are the fricative consonants /f/ and /s/ (as in *soup*), and /h/. When both of these mechanisms are used to create acoustic energy, the sounds are called *voiced fricatives* or *mixed-excited* sounds, as in /z/, /v/ and /zh/ (as in beige) sounds. *Plosives* are produced by a sudden release of compressed air. Examples of plosives are /p/, /t/, /k/, /b/, /d/, and /g/ (as in *good*). *Nasals* are produced when the nasal cavity is used as a

resonator, as in the /m/ and /n/ sounds in *man* and *no*.

A phoneme is the smallest unit of a word that when changed, changes the meaning of the word. For example, the words "cat" and "bat" each start with a different phoneme. At first glance, phonemes seem to describe specific speech sounds. However, they are a loose classification of more precise speech units called allophones, which account for physiological variations in the pronunciation of phonemes. For example the "t" sounds in "bat," "eighth," and "tree" are all pronounced slightly differently, and therefore are allophones of the phoneme /t/. Here, the variations are due to different places of contact of the tongue in the mouth. In *tea*, the tongue touches the alveolar ridge, which is the frontward part of the palate. This /t/ phoneme is thus an alveolar allophone. In *tree*, the tongue touches the palate a little further back, and is post-alveolar allophone; and in *eighth*, the tongue touches the teeth, so the allophone is dental. Allophones are indicated by the [ ] symbols.

Reflections of sound waves in the throat and mouth play an important role in the production of speech sounds. These reflections are caused by changes in cross-sectional area of the sound path. Although the largest of these changes occur at the vocal cords and at the lips, other cavities can also affect the reflections. These are the pharyngial cavity, the oral cavity, and the nasal cavity (see Figure 3). The pharyngial cavity and the oral cavity are collectively referred to as the *vocal tract*. The nasal cavity is coupled to the vocal tract by the trap door action of the *velum*, a soft piece of tissue at the back of the palate. The velum is also called the *soft palate*.

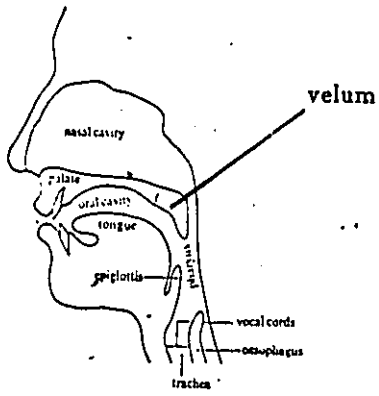

Figure 3: Cross-sectional view of the vocal mechanism (after O'Connor[6])

To understand how sound waves are reflected, consider the simplified model of the vocal tract shown in Figure 4:

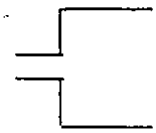

Figure 4: Simplified acoustic tube model of the larynx and the vocal tract

From transmission line theory, an increase in cross sectional area, and hence a decrease in acoustic impedance, produces a negative reflection coefficient; the polarity of the reflected wave is inverted. Conversely, a decrease in cross sectional area produces a positive reflection coefficient, and the polarity of the reflected wave is not inverted. A wavefront that originates at the larynx will undergo a negative reflection at the lips. This reflected wave, on coming back to the larynx,

will undergo a positive reflection, and head back towards the lips and so forth. This mechanism transforms a single impulse into an oscillating wave. If the acoustic tube consists of several sections of different cross-sections, such as is the case for the vocal tract, the reflections and resulting waveforms are more complex. The effect of the vocal tract on the shape of the acoustic waveform is seen in Figure 5. The top waveform is a 20 ms portion of the time-differenced waveform of a neutral vowel, and the lower trace is the same waveform with the effect of the vocal tract removed. A neutral vowel is one for which the major reflective boundaries are just the lip opening and the glottis, as shown in the simplified acoustic tube model of Figure 4.

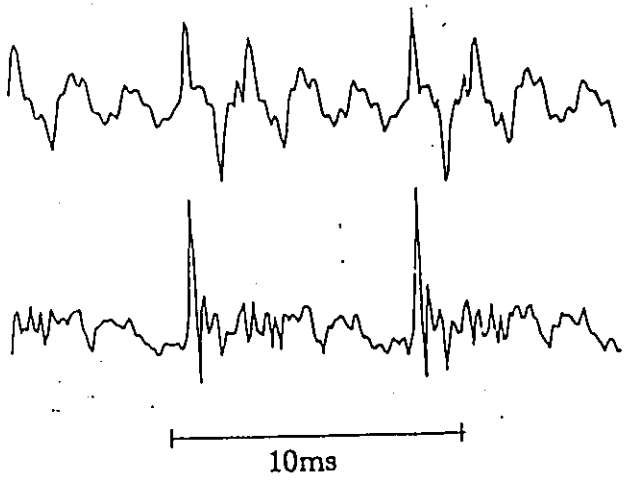

Figure 5: Effect of the vocal tract on the excitation (from [7]). The top waveform shows a time-differenced speech waveform, and the bottom window shows the same waveform with the effect of the vocal tract removed.

These reflections give rise to resonances called *formants*, which are characterised by their frequency and bandwidth. Human speech typically contains five formants in the 0-5 kHz range: the frequency of the lowest formant has an average

value of 500 Hz for male speech and varies from 300 Hz to 700 Hz; subsequent formants are spaced on average 1 kHz apart. In female speech, formant frequencies are on average 15% higher than in male speech, since the female vocal tract is typically 15% shorter. Formants produce peaks in the spectral envelope of the sound from the vocal cords, and it is mostly the frequencies of the formants that give these sounds their phonetic identity; there is a one-to-one correspondence between the identity of a steady vowel and the frequency of the first three formants.

The lip opening also affects the spectrum of speech sounds; from plane wave propagation theory, a wave passes more efficiently through an opening if its wavelength is much smaller than the opening. Hence, high-frequency components are transmitted more efficiently through the lips than low-frequency components. This effect, which tilts the speech spectrum by approximately 6 dB/octave, can be modelled as a time differentiation of the signal. The voiced speech spectrum therefore falls off at an approximate rate of -6 dB/octave. Figure 6 shows the log-amplitude spectrum of a segment of voiced speech, showing the -6 dB/octave tilt, as well as resonances caused by reflections of sound waves in the vocal tract.

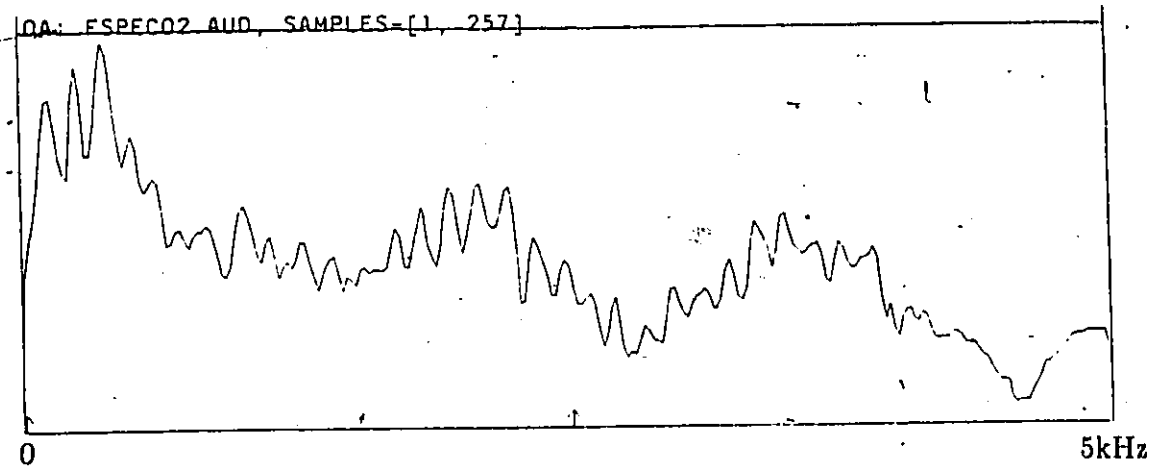

Figure 6: Short-time log-amplitude spectrum of voiced speech

In this thesis, the fundamental frequency is denoted by F_0, and F_x, B_x denotes the frequency and bandwidth of the xth formant, $x \in [1, ..., 5]$.

CHAPTER 2

VOCODER-BASED SPEECH OUTPUT SYSTEMS

2.1 Introduction

The preceding chapter gave introductory matter on speech sounds in general. In particular, it explained how the sound generated by the vocal cords is filtered by the vocal tract to produce the acoustic waveform that we call speech. This immediately suggests a linear-systems approach to speech synthesis, a principle known as the source/filter speech production model.

Vocoders — a contraction of *voice coders* — attempt to carry out a source/filter separation of speech sounds (see Figure 7). Separating the source and the filter allows the voice to be encoded efficiently, as both source and filter parameters change slowly in time relative to the rate of change of the acoustic signal. To carry out the source/filter separation, vocoders make the assumption that source and filter are independent. The assumption is not entirely true, since, for example, the muscular force needed to raise the pitch also raises the larynx; this shortens the pharynx and increases formant frequencies, allowing us to simulate pitch changes in whispered speech. Nevertheless, the independence assumption is sufficiently valid, since high quality speech can be synthesized from this principle. Moreover, because the ear is largely insensitive to phase, vocoders need only to determine the amplitude response of the vocal tract filter *i.e.*, the

envelope of the amplitude spectrum of the speech waveform. As for the source, it is often represented as a single parameter such as the pitch of the driving function in the case of voiced sounds together with a measure of loudness. Several methods can be used to determine the pitch[8][9], such as an autocorrelation method on the speech signal, or the use of an electroglottograph, an electronic device which monitors the activity of the vocal cords[10].

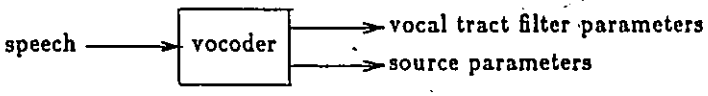

Figure 7: Source/filter separation of speech sounds.

Vocoders are classified according to how they how they estimate the frequency response of the vocal tract filter:

2.2 Channel vocoders

Channel vocoders[11] feed the speech signal to a bank of bandpass filters and measure the short term average power output of individual filters. The encoded parameters are the bandpass powers. These vocoders must have enough channels to get a good approximation of the spectrum envelope, especially around the formant peaks. Channel vocoders typically have 15 to 20 channels, and use either analog filters, digital filters, or FFT's to obtain bandpass powers. The channels are spaced farther apart at higher frequencies to reflect the frequency resolution of the ear[12].

2.3 LPC vocoders:

LPC vocoders[13] model speech as the output of an all-pole filter *i.e.*, resonant filters connected in series. Under this assumption, a technique called *linear prediction* (LP)[23] is used to analyse the resonances of the vocal tract, and the encoded information consists of parameters that describe an all-pole model of the vocal tract filter. LPC fits the spectrum of an all-pole filter to the spectrum envelope of the speech segment being analysed. Even when the all-pole assumption is invalid, such as during nasal sounds, LPC still does a reasonable job of estimating the spectral envelope of the speech. The spectrum fitting is not done on the usual least-squares criterion, but on a criterion that concentrates on a fitting of the high energy regions of the spectrum at the expense of weaker regions. This is because LPC does a least-squares fit to the waveform, and high energy parts of the signal affect the spectrum the most. Linear prediction is discussed at greater length in Section 2.7.

2.4 Formant vocoders

Formant vocoders[14][15] do a formant analysis using frequency domain methods. The Fourier transform of the speech waveform is taken, and special algorithms determine formant parameters. Once the formants are determined, the synthesis is produced using either a cascade or parallel filter configuration. The cascade configuration is less flexible because it matches spectra that fit an all-pole model *i.e.*, resonances in series. The parallel configuration allows for deviations from

the all-pole speech production model because resonances are specified by frequency, amplitude and bandwidth, as opposed to frequency and bandwidth only.

2.5 Vocoder applications

The source/filter separation performed by vocoders leads to an efficient coding of the speech, because both the source and the filter change slowly with time. This makes vocoders good candidates for systems concerned with minimizing data rates (for speech transmission) or memory requirements (for speech storage). Once the filter parameters are determined, speech can be synthesized by exciting the filter with either periodic impulses or white noise, depending on whether the sounds are voiced or voiceless.

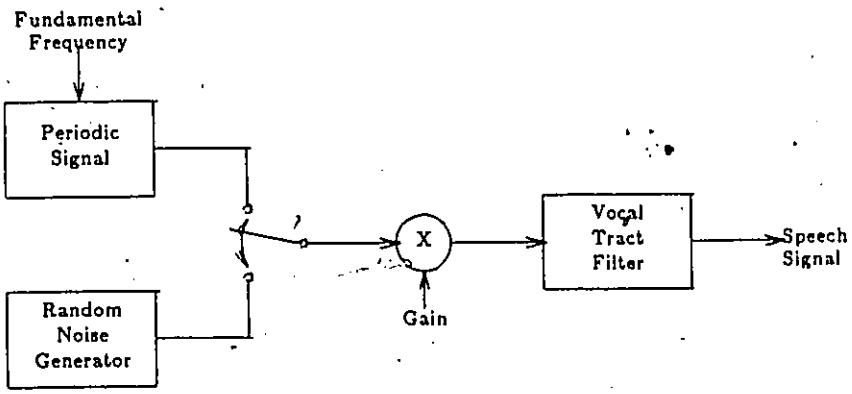

Figure 8: Speech production model

A source/filter separation allows for easy modifications of the speech, such as speed and voice quality changes. This flexibility is useful for text-to-speech systems, because it enables them to change timing and intonation patterns

(prosody). Vocoders are also used for research in speech production and speech perception[3].

2.6 Alternatives for speech output

2.6.1 Analog waveform storage: This technique simply stores the waveform without any attempt to compress it. Various media are used to store the speech waveform: voice clocks use glass disks; other systems, such as telephone answering machines, magnetic tapes. With waveform storage; the quality of the output message is high, but the output is fixed; messages must be replayed in their original form. Another problem arises when new words need to be added to an existing system: new recordings must be made, and if the original speaker cannot be found, the entire material must be re-recorded and re-edited with a new speaker.

2.6.2 Waveform coders: Waveform coders store digitally encoded speech waveforms. Waveform coding techniques compromise data rates, equipment complexity and speech quality. Just as with analog waveform storage, waveform coders replay messages in their original form, and the original speaker must be found if additional speech needs to be added to the system.

- *Pulse Code Modulation (PCM)*[16] is digital waveform storage. With PCM, the speech waveform is sampled and sample values are digitized. The hardware is simple and, provided the bit rate is high enough, the quality is excellent. The drawbacks of PCM reside in the high memory requirements to store a large vocabulary, and in the fixed output. In an

effort to alleviate both problems, waveform storage systems usually concatenate isolated words to construct messages. However, pitch, timing, and intonation patterns are often wrong, making it difficult to understand these messages.

- *Delta-modulation*[17] transmits differences in sample amplitudes as opposed to absolute sample values. The simplest delta-modulation scheme transmits a single bit per sample; this corresponds to the decision of adding or subtracting one quantization level to the signal being reconstructed. A danger with this scheme is *slope overload*, where the input signal varies too fast to be tracked by the encoder. This problem is avoided by increasing the sampling frequency, but this is not an efficient solution. Another solution is to replace the one-bit quantizer with a PCM coder to encode sample-to-sample differences. This arrangement is known as *differential PCM (DPCM).* Another scheme, adaptive diffenrential PCM (ADPCM)[18], adaptively adjusts the size of the quantization steps to increase the dynamic range.

2.6.3 Parametric representations: These methods encode the speech waveform in terms of parameters that describe a model for speech production. The LPC method described earlier belongs to this class. *Adaptive Predictive Coding (APC)* is a similar method that tries to encode a waveform by adapting coefficients of a predicting filter until an error criterion is met. The transmitter performs a "dummy synthesis" and uses the error signal, the difference between the synthetic and the original waveform, to design the predicting filter. To

reconstruct the speech, the error signal must be available at the receiver, and so it must be encoded and transmitted along with the filter coefficients. The decoding error on the error signal is compensated by including a coder-decoder block in the transmitter's predictor feedback loop.

Due to its iterative nature, APC is computationally more expensive than straight LPC, which derives the filter coefficients in a direct manner. The advantages of parametric representations are that storage requirements are cut down, and that speech sounds can be modified by altering the value of the parameters. This can be helpful to smooth transitions at boundaries between concatenated speech sounds, and to change pitch and timing patterns. Parametric representation techniques, however, are computationaly more expensive than waveform coding techniques, and the quality of the output is generally poorer.

2.6.4 Text-to-speech systems: Text-to-Speech systems[20] convert written text into an acoustic signal. The first step is usually to convert the orthographic representation into a phonemic transcription, which typically includes the name of phonetic segments, stress marks, pitch, and timing information. The second step is to convert the phonemic transcription into sound. Several techniques can be used: demisyllable based systems[19] concatenate demisyllables. For example, the word "peak" may be synthesized by concatenating the acoustic waveforms of "pee," and "eek."

With the so-called rule based approach[20][21] phonetic transcriptions are converted into formant trajectories. The conversion from formant tracks to

acoustic signal can be achieved with a digital filter or an analog filter that is digitally controlled.

A hard task to achieve with text-to-speech systems is to get correct timing and intonation patterns within a sentence, such as in "The *cat* ate the mouse" (emphasis on subject) and "The cat *ate* the mouse" (emphasis on verb). In this example, the proper intonation is deduced by an understanding of the preceding text. Sometimes, however, the intonation can be deduced by analyzing the grammatical structure of the sentence, such as in "Stand up!" Paul said.

Text-to-speech systems enjoy a very low bit-rate (< 100 bits/sec) at the expense of substantial computation requirements. An advantage of text-to-speech systems over other speech output systems is that any text can be converted into speech without adding anything new to a system. Thus, changing the speech output is as easy as editing a text file.

2.7 Speech analysis using LPC

This section presents more details on LPC. This information is necessary to understand specific aspects of the LPC vocoder developped at the NRC, and to understand the improvements made by the author.

One of the most common speech analysis methods during the past twenty years has been LPC. LPC is particularly predominant in digital speech communication systems because it provides low bit-rate transmission without excessive loss of inteligibility. Several American and Japanese companies[22]

currently manufacture special-purpose chips for LPC synthesis, and as a result, these chips are widely available and inexpensive.

The basic principle behind LPC analysis is that the vocal tract can be modelled as an all-pole filter. However, during speech, the shape of the vocal tract is continously changing as a function of time. To deal with this, an LPC analysis involves time-windowing the speech into analysis segments; assuming the shape of the vocal tract is constant for the duration of a segment, the speech waveform in each segment is modelled as the impulse response of an all-pole filter.

The term "linear prediction" (LP) refers to a mathematical method that is used to compute the parameters of an all-pole filter from autocorrelation properties of its impulse response. Provided that the waveform being analysed actually is the impulse response from an all-pole filter, that the analysis order corresponds to the number of resonances present, and that the waveform being analyzed contains only freely decaying oscillations *i.e.*, if the instant of excitation is excluded, the LP analysis will determine the all-pole filter exactly.

Consider the following all-pole production model of a single speech segment:

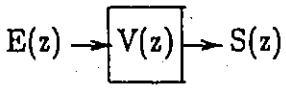

Figure 9: All-pole speech production model

In Figure 9, $E(z)$ represents the excitation function, assumed to be spectrally flat, $V(z)$ is an all-pole filter that represents the vocal tract, and $S(z)$ represents a speech segment. LPC actually tries to flattens the amplitude spectrum of the signal it analizes by constructing an all-zero filter $A(z)$ that will cancel the poles of $V(z)$. The analysis model is thus

$$E(z) = S(z)A(z), \tag{1}$$

with

$$A(Z) = \frac{1}{V(z)}. \tag{2}$$

The synthesis equation is then

$$S(z) = E(z)\frac{1}{A(z)}, \tag{3}$$

in which $S(z)$ is modeled as the impulse response of an all-pole filter. Both of these processes are depicted in Figure 10. Shown are log-magnitude (LM) spectra (top) and idealized time waveforms (bottom). The left of the figure shows a speech signal that passes through the filter $A(z)$. This produces and impulse-like error signal since all the predictibility of the speech signal has been removed. On the right side, the error signal is fed to the filter $\frac{1}{A(z)}$ to produce an exact replica of the original speech signal.

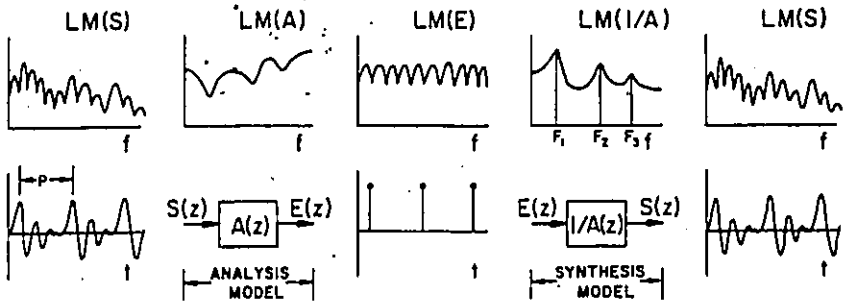

Figure 10: Analysis and synthesis models for a voiced sound (from [23]). Shown are log-magnitude (LM) spectra (top) and idealized time waveforms (bottom). The left of the figure shows a speech signal that passes through the filter $A(z)$. This produces and impulse-like error signal since all the predictibility of the speech signal has been removed. On the right side, the error signal is fed to $\frac{1}{A(z)}$ to produce an exact replica of the original speech signal.

Re-writing the analysis equation,

$$E(z) = S(z)\hat{A}(z),$$

in which

$$A(z) \equiv \sum_{i=0}^{M} a_i z^{-i} \qquad (a_0 = 1). \tag{4}$$

In the sampled data domain, this becomes

$$e(n) = \sum_{i=0}^{M} a_i s(n-i) = s(n) + \sum_{i=1}^{M} a_i s(n-i)\,. \tag{5}$$

If we define the predicted speech $\hat{s}(n)$ as

$$\hat{s}(n) \equiv -\sum_{i=1}^{M} a_i s(n-i), \tag{6}$$

then

$$e(n) = s(n) - \hat{s}(n), \tag{7}$$

where $e(n)$ is seen as the prediction error. The objective of linear predictive analysis is to minimize the square of the error between the actual speech and the predicted speech. The total square error α is given by

$$\alpha = \sum_{n=n_0}^{n_1} e^2(n) = \sum_{n=n_0}^{n_1} \left[\sum_{i=0}^{M} a_i s(n-i) \right]^2 = \sum_{n=n_0}^{n_1} \sum_{i=0}^{M} \sum_{j=0}^{M} a_i s(n-i) s(n-j) a_j, \tag{8}$$

where $[n_0, n_1]$ is the interval over which the prediction error is minimized. Defining

$$c_{ij} \equiv \sum_{n=n_0}^{n_1} s(n-i) s(n-j), \tag{9}$$

the total squared error is

$$\alpha = \sum_{i=0}^{M} \sum_{j=0}^{M} a_i c_{ij} a_j . \tag{10}$$

Minimization of α is obtained by setting

$$\frac{\partial \alpha}{\partial a_j} = 2 \sum_{i=0}^{M} a_i c_{ij} = 0 \qquad \text{for } j = 0, 1, 2, \ldots, M .$$

Since $a_0 = 1$, the above equation yields

$$\sum_{i=1}^{M} a_i c_{ij} = -c_{0j} \qquad j = 0,1, 2, \ldots, M \tag{11}$$

These M equations must be solved to minimize the prediction error. The solution

of the above set of equations contains two steps. The first step is to compute the c_{ij} terms, or C matrix. The second step is to solve simultaneously M linear equations.

Two methods are commonly used to compute the C matrices: the *covariance* method and the *autocorrelation* method. The covariance method produces symmetric C matrices that are solved efficiently with Cholesky's recursive decomposition method[24]. The autocorrelation method produces a C matrix whose elements depend only on the distance from the main diagonal. This type of matrix is called a Toeplitz matrix and is resolved even more efficiently by Levinson's method[23], which is also a recursive method. The autocorrelation method is simpler than the covariance method, and is generally considered less accurate. However, the accuracy issue is still contentious, as a colleague has shown[25]. The autocorrelation method has the added advantage that it has a built-in stability test to tell whether the next recursion step will result in an unstable filter. Details on both methods is presented in section 3.7.

Figure 11 shows a block diagram of an LPC analysis.

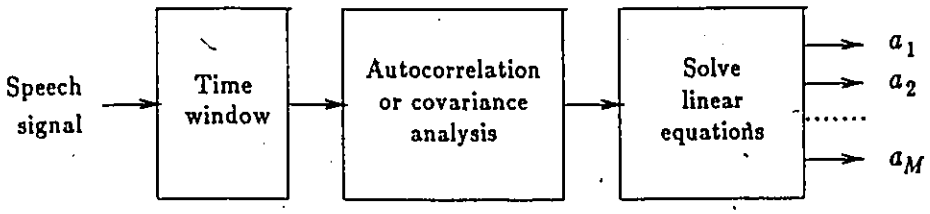

Figure 11: Linear Prediction analysis

An important point is that a suitable number of LPC coefficients (~ 10) permits us to obtain a good approximation to the formant frequencies and bandwidths of each speech segment being analyzed. This can be seen in the relation

$$\frac{S(z)}{E(z)} = \frac{1}{A(z)} = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots} = \frac{1}{(1 - \frac{z_1}{z})(1 - \frac{z_2}{z}) \cdots (1 - \frac{z_M}{z})}$$

where the poles $(z_1, z_2, \ldots, z_M)$ correspond to the resonances of the speech waveform. Figure 12 shows the amplitude spectrum of a voiced speech segment together with the spectral envelope defined by the filter equation $\frac{S(z)}{E(z)}$.

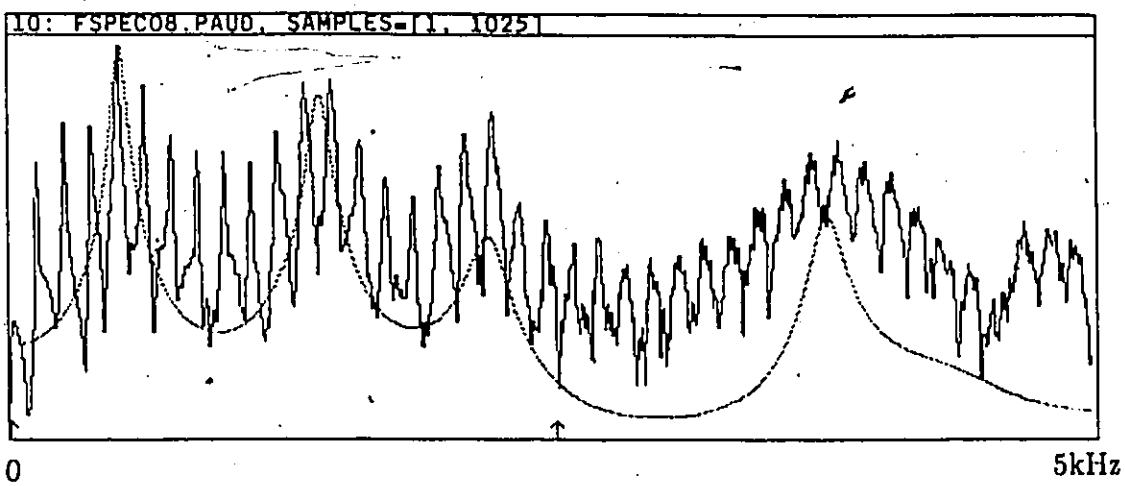

Figure 12: Short time amplitude spectrum and linear prediction envelope (smooth curve) of a voiced speech segment. The spectral envelope estimated by the LPC analysis is usually a good approximation to the short-time spectral envelope of the speech segment, and its peaks generally correspond to formants.

The M poles $(z_1, z_2, \cdots, z_M)$ usually consist of $\frac{M}{2}$ complex conjugate

pole pairs, although sometimes real poles are found. The complex conjugate pole pairs, commonly visualized as cartesian coordinates in the z-plane ($z_n = Re(z) \pm jIm(z)$), may be converted into polar coodinates (θ, $|z|$) in the same plane, with the relationship

$$\theta = arg(z)$$
$$|z| = \sqrt{zz^*} = \sqrt{(Re\ z)^2 + (Im\ z)^2}$$

where θ is the angular position of the pole, circle, and $|z|$ is its radial distance from the origin. The polar coordinates may be converted into formant frequencies and bandwidths if desired. The frequency of a pole is given by

$$\frac{\theta}{2\pi} f_s \ [Hz] \tag{12}$$

where f_s is the sampling frequency. The bandwidth of a pole is a measurement of its radial distance from the unit circle, and is usually defined as the point where the power is 50 % of the power at resonance. In this case, the bandwidth is

$$\frac{-f_s}{2\pi} \log |z|^2 \ [Hz], \quad 0 \leq |z| \leq 1 \tag{13}$$

where $|z|$ is the argument of the pole. If $|z|^2$ is equal to one, *i.e.*, the pole is on the unit circle, and the bandwidth is zero. The closer the pole is to the origin, the larger the bandwidth is, and poles outside the unit circle correspond to unstable filters.

Another important point, especially for real-time communications, is

that LPC produces a set of linear equations that can be solved directly. This is not the case when the speech is modelled as the response of a pole-zero filter; least-squares estimation then produces non-linear equations, even in the simplest cases. An iterative method must then be used, and the method is not guaranteed to converge to a solution. The all-pole model of the vocal tract is based on the following assumptions:

- The transverse dimensions of the vocal tract are small with respect to the sound wavelength, so that sound propagation through the vocal tract can be treated as a plane wave. (This assumption is valid for frequencies up to around 5 kHz).
- Losses due to wall vibrations are negligeable.
- The vocal tract behaves as an unbranched tube: *i.e.*, the path through the nasal tract is closed off.

Another important point is that the LPC analysis is biased towards matching the high energy parts of the spectrum. As well, the error signal (the residual) can be used to obtain the glottal air flow waveform[5], as it represents a speech waveform from which the effect of the vocal tract has been removed.

2.8 Details on autocorrelation and covariance analyses

Both of these methods are techniques used to estimate autocorrelation properties of the speech signal.

2.8.1 The covariance method:

The elements of the C matrix are computed from Eq. (9):

$$c_{ij} = \sum_{n=M}^{N-1} s(n-i)s(n-j), \quad i, j = 0, 1, 2, \cdots, M$$

where N is the number of speech samples and M is the order of the analysis. From this equation, we see that $(N-M)$ terms are used to compute individual c_{ij}'s. Thus, we must have the condition that $N > M$, and so we need at least $M+1$ speech samples in the window. Figure 13 shows how these parameters relate to a speech frame for the covariance analysis.

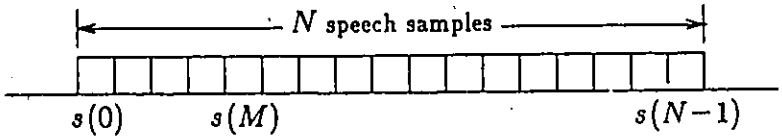

Figure 13: Covariance window

Windows are typically 25 ms long in order to average over several excitation periods, and are usually overlapped by a constant amount.

For the covariance method,

- all elements of the C are computed from an equal number of terms;
- the error is minimized over the $[M, N-1]$ interval;
- the C matrix is symmetric i.e. $c_{ij} = c_{ji}$.

The error is given by

$$e(n) = \sum_{i=0}^{M} a_i s(n-i) \quad \text{for } n = M, M+1,..., N-1 . \tag{15}$$

2.8.2 The autocorrelation method: This method assumes that $s(n) = 0$ outside the window.

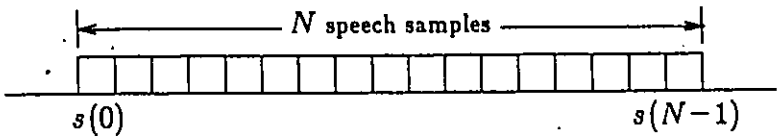

Figure 14: Autocorrelation window

$$c_{ij} = \sum_{n=n_0}^{n_1} s(n-i)s(n-j) = \sum_{n=-\infty}^{\infty} s(n-i)s(n-j) = \sum_{n=-\infty}^{\infty} s(n)s(n+|i-j|)$$
$$= \sum_{n=0}^{N-1-|i-j|} s(n)s(n+|i-j|) \equiv r(|i-j|) \equiv r(j),\ j = 1, 2, 3,..., M \tag{16}$$

Again we must have that $N-M > 0$. For the autocorrelation method,

- the analysis window is usually long enough to contain several glottal cycles (typically 3 to 5).
- the number of non-zero terms used to compute the c_{ij} terms decreases as $|\ i-j\ |$ increases, so the $r(j)$ coefficients are biased.
- the C matrix is Toeplitż i.e. $c_{ij} = r_{|j-i|}$

The error is given by

$$e(n) = \sum_{i=0}^{M} a_i s(n-i) \quad \text{for } n = 0, 1,..., N+M-1 \tag{17}$$

Either of these methods produces linear equations that, when solved, give the

coefficients a_i of an inverse filter $A(z)$. It can easily be shown that the DFT of the a_i's is the spectrum of the inverse filter.

2.9 Standard LPC vocoders

This section describes the steps involved in a standard LPC analysis. Variations on the standard design are presented in the next section.

Although considerable research has been made on LPC vocoders for real-time speech communications, relatively little work has been done specifically for speech storage, where the quality of the speech is a major concern, and where LPC encoding does not need to be performed in real-time. For this reason, "standard LPC vocoders" are those used in real-time speech communication systems.

The first process in a LPC vocoder is the segmentation of the speech signal. The analysis is said to be *pitch-synchronous* if segments are synchronized with vocal cord periods (glottal cycles). When this is not the case, the analysis is *pitch-asynchronous*. In this case, the speech is divided into frames (segments) of equal length, usually 20 to 50 ms.

The next step is a preemphasis operation on the speech signal. The main advantage of preemphasis is that it makes the LPC fitting process pay equal attention to fitting all parts of the spectrum. As noted in Section 2.1, the spectrum of the excitation signal from the vocal cords is smooth and tilted at -12 dB/octave. After passing through the vocal tract, the amplitude spectrum

contains peaks that correspond to formants, but still has a general -12 dB/octave tilt. The lip radiation effect gives this spectrum envelope a +6 dB/octave tilt, so that the spectrum of the speech we hear has a resulting -6 dB/octave tilt. Preemphasis is done by filtering the speech with the transfer function

$$H(z) = 1-\beta z^{-1}$$

where β is called the *leak factor*. This transfer function performs a sample-to-sample differentiation. If the leak factor is close to unity, H(z) approximates best a true time differencing operation, which tilts the spectrum of a signal by +6 dB/octave. Choosing a leak factor close to unity (0.9 to 0.99) also minimizes non-linearities in the phase spectrum. Figure 15 shows the amplitude and phase spectrum of this function for different values of β.

Because the spectral characteristics of the excitation from the vocal cords do not vary widely, the fixed preemphasis will always approximately flatten the spectrum of voiced speech. Voiceless sounds usually have a flat amplitude spectrum and are therefore inappropriately preemphasized, but since an accurate analysis of resonances is less important here, the problem is not serious.

Next, given a speech segment, linear equations are computed using either the *autocorrelation* or *covariance* analysis method. These linear equations, when solved, provide parameters that fully describe the all-pole prediction filter. The transmitted information (the encoded speech) consists of these parameters

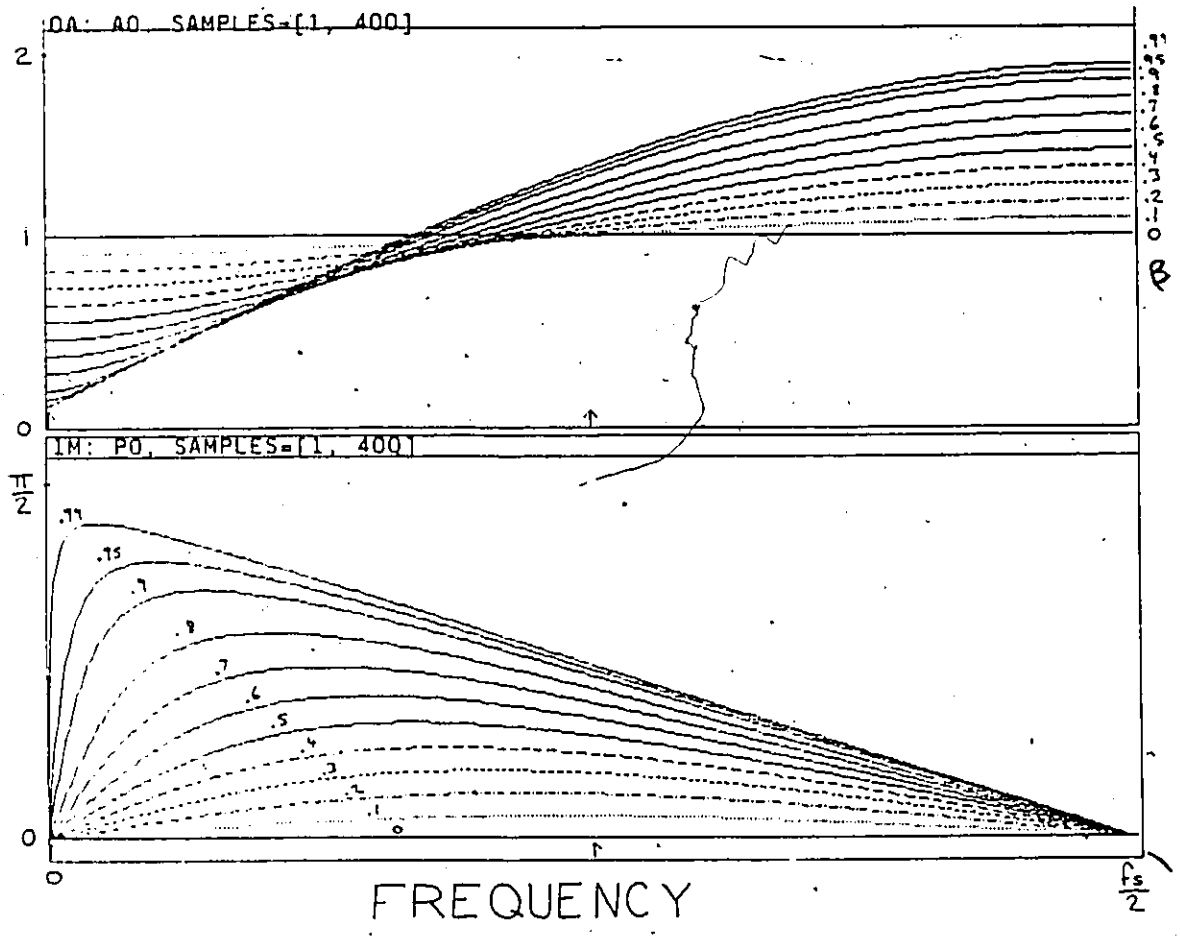

Figure 15: Amplitude spectrum (top) and phase spectrum (bottom) of preemphasis filter.

along with parameters which describe the excitation source (either a periodic signal or white noise), along with a measure of its power.

2.10 Other LPC vocoders:

LPC vocoders are typically used to compress speech for real-time digital transmission purposes; they offer bit-rate reductions more than an order of

magnitude compared to standard PCM. Probably motivated by practical concerns, work in this area has been considerable, and several kinds of LPC vocoder systems have been developed; they differ mostly in the way they excite the vocal tract filter. If the residual signal from the LPC analysis is used, the reconstructed speech will be exact (see Figure 10). An uncompressed residual, however, offers no data rate reduction. Hence, most kinds of vocoders variants around differ in how they encode the residual. Some kinds, because of their importance, have been given specific names:

- *Residual-Excited LPC (RELPC)*[26] encodes the residual before it is transmitted. The objective is to encode the residual in a way that preserves its perceptually important features. An example of a residual encoding scheme is to low-pass filter the residual and transmit it at a reduced data rate. The receiver artificially restores high frequency components of the residual by methods such as spectral aliasing.

- *Multipulse LPC*[27], as its name suggests, uses a series of impulses to excite the vocal tract filter during a single glottal cycle. There is a main pulse that corresponds to the main excitation, and a series of smaller pulses to correct inadequacies of the LPC filter. It is the amplitude and position of these pulses that is encoded for transmission. Typically, 10 - 20 pulses are used for every 20 ms of speech.

- *Code-excited linear prediction (CELP)* vector-quantizes glottal cycle residuals. The choice of a vector for a particular frame is made by exhaus-

tively trying out all possibilities and choosing the vector that produces the best synthesis. Typically, a 10 bit code is used to encode 1024 residuals. The spectral coefficients can also be vector-quantized.

CHAPTER 3

LPC AT THE NRC

3.1 Introduction

This chapter presents specific information about the pitch-synchronous LPC vocoder developed at the NRC. The purpose of this system is to provide speech of high quality and high flexibility for aeronautical applications, and it is stressed that the requirements of such a system are different from those involved in real-time communications systems:

- there is no need for real-time LPC analysis
- the encoded speech is stored in system memory.
- encoding of the speech waveform is necessary not because of bandwidth limitations, but for efficient speech storage.

3.2 Functional description

The LPC speech analysis and synthesis system consists of eight programs, which are mainly written in "C" and run under the UNIX operating system. The system follows the UNIX philosophy of simple modular units, where the output of one program may be the input to the next.

As we saw earlier, the function of a vocoder is to perform a source/filter

separation of a speech waveform. The process consists of analyzing a digitized speech waveform, and to produce a file that contains parametric information on the source and on the vocal tract filter. At the NRC, the source information consists of the type of source used (periodic signal or white noise, for voiced or voiceless frames respectively), and of a gain factor. Information on the filter consists of parameters that describe the all-pole filter from the LPC analysis. The NRC vocoder encodes these parameters into formant frequency-bandwidth pairs. Both source and filter parameters are computed on a frame-to-frame basis, and they are neatly stored in vector form, one vector per frame. Global modifications, such as doubling the pitch-period, or increasing a formant frequency by 10%, are thus easily achieved. Moreover, this parameter file represents a highly compressed version of the original speech signal, as it contains all the information necessary to reconstruct the acoustic waveform.

3.3 System details

Speech spoken in an anechoic chamber together with the signal from a laryngograph (or electroglottograph) is digitized with no aliasing and no phase distortion. The laryngograph measures the r.f. impedance between two electrodes placed across the larynx, which is proportional to the area of contact of the vocal cords. Both channels are digitally encoded at 44 kHz and recorded on a video tape using a Sony F1 PCM recorder. For input to the computer, they are low-pass filtered at 7 kHz, sampled at 20000 sps, sharply low-pass filtered at 5.0 kHz, and finally down-sampled to 10000 sps. The 5 kHz low-pass filtering is done

twice, forwards and backwards, to double the cutoff rate and to remove phase distortions. The effective cutoff rate is -800 dB/decade, and the cutoff frequency is 5.0 kHz. Figure 16 shows both of these signals for the utterance of the word *zero*.

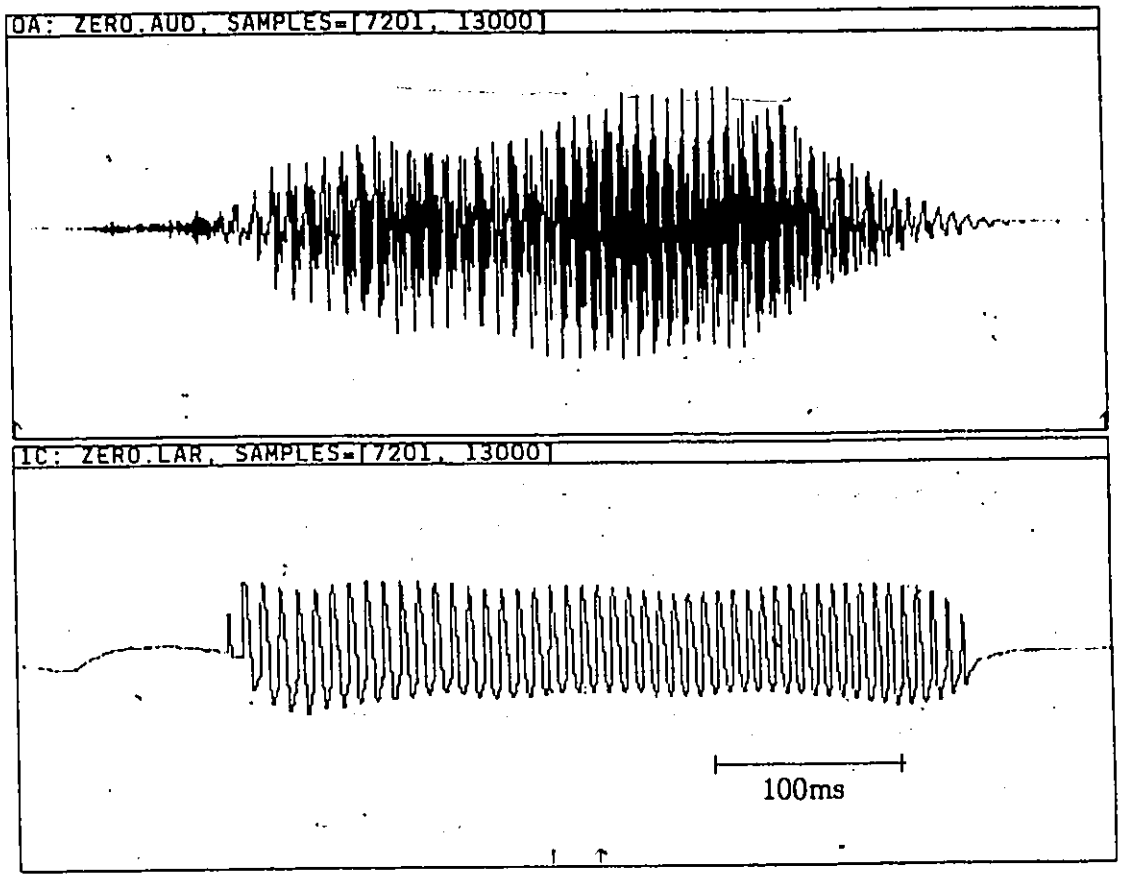

Figure 16: Audio signal (top) and laryngograph signal (bottom) for the utterance of the word "zero."

Research for this thesis has shown that as far as the quality of the output synthesis is concerned, the high cutoff rate and linear phase is not necessary, as low-pass filtering the 20000 sps speech once with a 6th order Butterworth filter,

down-sampling to 10 kHz and re-doing the analysis produces no perceptible degradations in the synthesis.

If the analysis is to reflect accurately the true formant frequencies and bandwidths of the vocal tract, the recordings should not contain reverberation. However, research for this thesis has shown that reverberation-free recordings are not necessary for a high quality speech output. A small amount of reverberation is unnoticeable as far as synthesis quality is concerned, though large amounts of reverberation produces a synthesis that sounds reverberant. In either case, the formant frequencies and bandwidth values are not those of the vocal tract, but those of the combined vocal tract and the room in which the recordings are made. Hence, if the objective is to estimate accurately the vocal tract parameters, the recordings should be of anechoic-chamber quality. As well, comparison between analyses with and without reverberation shows that formant bandwidths are more susceptible to reverberation than formant frequencies. Figure 17 illustrates the effect reverberation has on F_1 and B_1 for the sentence "We were away a year ago." The top window shows how F_1 varies as a function of time, and the bottom window shows how B_1 varies as a function of time. In each window, the solid line is the parameter value without reverberation, and the dotted line is the parameter value with reverberation. The amount and type of reverberation is typical of that of a 15 feet room (two opposing walls 15 feet apart).

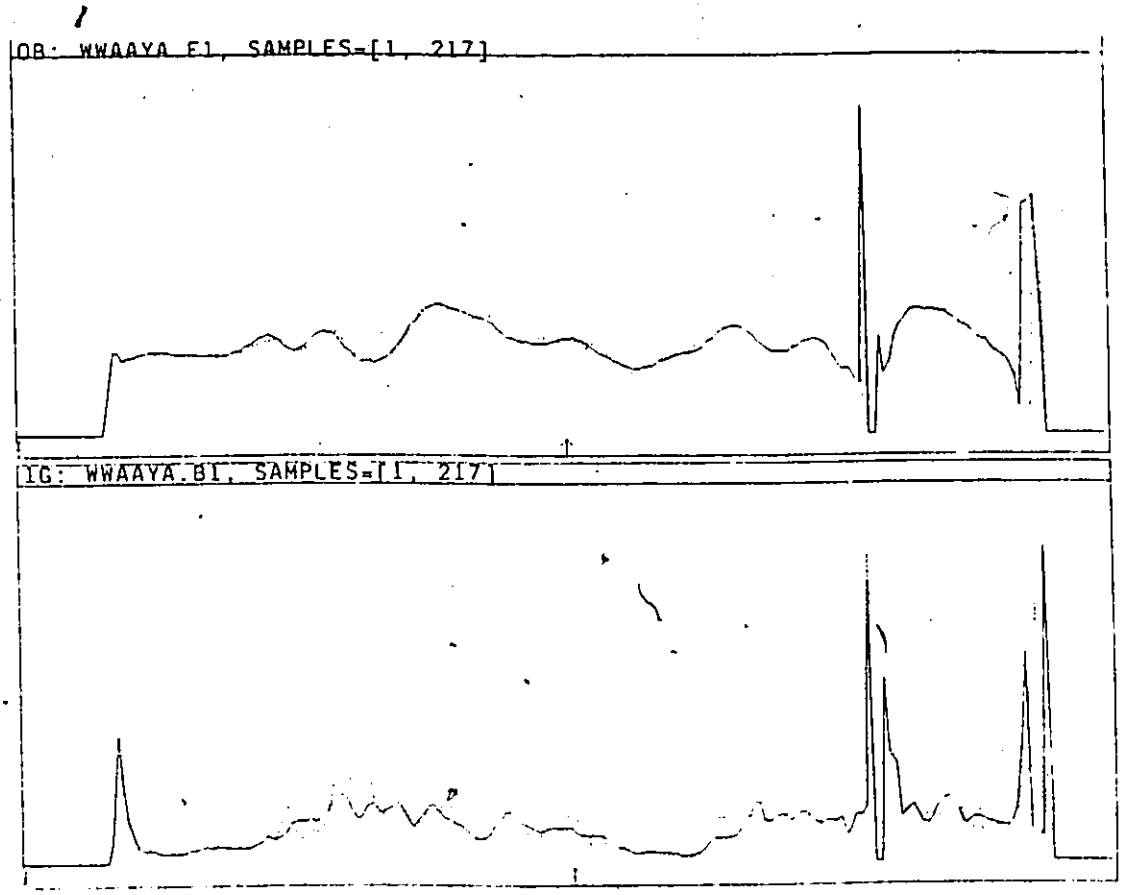

Figure 17: Effect of reverberation on F_1 (top window) and B_1 (bottom window). Both windows correspond to the sentence "We were away a year ago." In each window, the solid line is the parameter value without reverberation, and the dotted line is the parameter value with reverberation.

The digitized laryngograph signal is differentiated, and the impulses in this signal are taken to correspond to instants of glottal closure, thus providing an indication of voicing and fundamental frequency F_0 (see Figure 18).

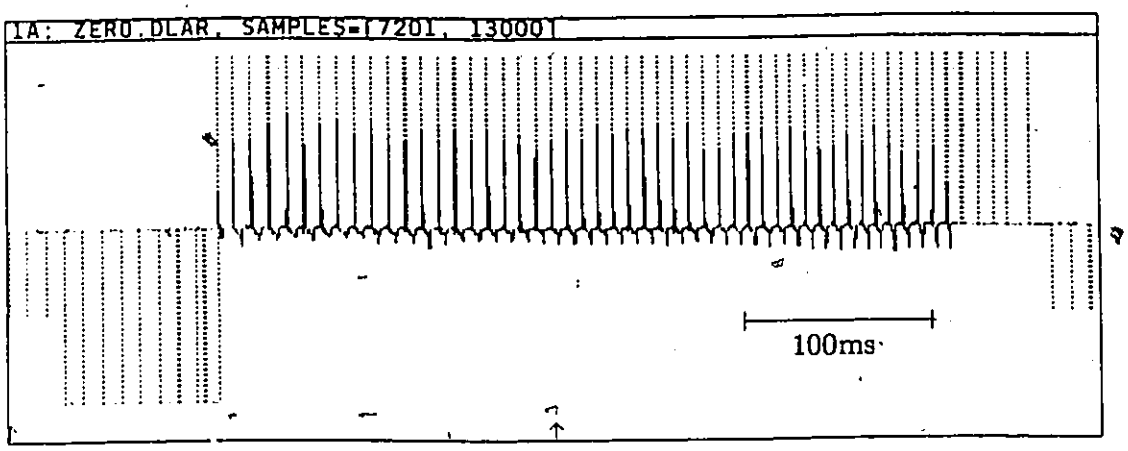

Figure 18: Differentiated laryngograph signal with overlaid segment boundaries (dotted vertical lines), as determined by the segmentation algorithm. Dotted lines going to the top of the window indicate voiced frames, to the bottom of the window, voiceless, and half-way down, silent.

The speech segmentation program detects these peaks, and produces a file of two-element records (Figure 19). The first element, called the *frame size*, controls the rate of change of the vocal tract as a function of time: the number is the time (in number of samples) during which the filter coefficients are constant. The second element indicates the frame type: a number greater than zero indicates a voiced frame (pitch-period times ten), "-1" indicates a voiceless frame, and "0," a silent frame. If a pitch-period is not found, the frame size is set at 100 samples (10 ms at 10000 sps). The pitch-period times ten is used because the speech segmentation program estimates the pitch-period to a tenth of a sample. It does this by fitting a quadratic curve to the three samples around a peak of the differentiated laryngograph signal. Although voiceless frames are encoded as -1, subsequent programs interpret any negative number in this field as a voiceless

frame. This classification system was chosen to ease voice modification. For example, to double the pitch-period, the entire second column can by multiplied by a factor of two, and this will not change the frame types.

```
100    0
100    0
 90    0
100   -1
100   -1
 78   -1
 78  780
 98  981
 94  941
```

Figure 19: Sample output from the speech segmentation program. The first column is the frame size (in number of samples of the audio file), and the second column indicates the frame type. A "0" indicates a silent frame, "-1" (or any negative number) indicates a voiceless frame, and a positive number represents the pitch-period times ten of a voiced frame.

Now that speech frames are defined and classified, the audio signal is preemphasized for analysis. Contrary to other vocoders, only the voiced frames are preemphasized. The preemphasis function is

$$H(z) = 1 - .95z^{-1}$$

The next step is the LPC analysis *per se*. For both voiced and voiceless speech, a 10th order covariance-method LPC analysis is then applied to rectangular analysis frames beginning up to ten samples before glottal closure and ending just before the next glottal closure. The analysis frames for voiceless speech are set at 10 ms.

Covariance matrices are smoothed with a quarter-half-quarter operation: the final value of a particular matrix element is 1/4 the values of the same element from the previous matrix, plus 1/2 the value of the element in the current matrix, plus 1/4 the value of the same element from the following matrix.

Any instabilities in the filter specified by the covariance-method analysis are corrected by solving the predictor polynomial and reflecting the unstable poles inside the unit circle. Complex-conjugate pole pairs are converted to frequency and bandwidth pairs, generally corresponding to formants. Real poles are encoded as negative numbers.

The last frame parameter computed is the gain factor for the excitation. This gain is the scaling factor that must be applied to the excitation signal (the source) in order to have the power in the synthetic signal match the power in the audio file. The power in the synthetic waveform depends on two things: the power of the excitation function and on the ringing from previous frames. To find the gain factor, the filter memory is cleared and the filter is excited with an excitation of unit power; this produces a vector **u**. Next, the filter memory is loaded with the values from the previous frame, which contains the ringing from the previous cycles. The excitation is set to zero and the filter is allowed to "ring," producing a vector **q**. The gain g is then obtained by solving the quadratic equation

$$\langle \mathbf{s}^2 \rangle = \langle g\mathbf{u} + \mathbf{q} \rangle^2$$

The encoded value for the gain is proportional to the log power of the scaled

excitation. Figure 20 shows a sample output of the complete analysis for eight consecutive frames.

fs	pp	pow	f1	b1	f2	b2	f3	b3	f4	b4	f5	b5
75	746	8430	359	85	1004	567	2305	471	3415	565	4428	580
86	863	7797	332	46	612	118	2237	72	3603	206	4072	197
86	862	8682	337	27	678	100	2201	65	3517	101	3959	79
86	857	9239	339	17	759	81	2191	74	3502	63	3907	47
83	831	9999	345	14	892	68	2178	85	3494	57	3893	41

Figure 20: Five output records from the LP analysis. "fs" is the frame size, "pp" stands for pitch-period, "pow" is the logarithmically encoded excitation power, and the rest of the parameters are the pole frequencies and pole bandwidths of the all-pole model of the vocal tract filter.

The "voice coding" is now completed. To produce synthetic speech, records from the parameter file are read one at a time. The first parameter, the frame size, indicates how many samples are to be written out to the output file before the filter coefficients are updated. If the second parameter is "0", the frame is silent, and samples of value "0" are written out to the output file. If the second parameter is not zero, then the frame is voiced or voiceless. In any case, pole frequency-bandwidth pairs are reconverted to predictor coefficients, the filter taps are updated, and the filter is excited by a scaled driving function (noise or impulse-like function). The filter delay line is cleared when going into or out of a voiced region, and the filter coefficients are updated only at the beginning of a pitch-period (pitch-synchronous synthesis). Figure 21 shows the synthesis filter. The output of this filter needs to be deemphasized to compensate for the preemphasis that was carried out on the input speech.

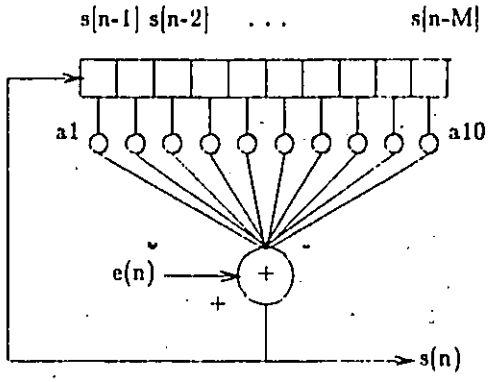

Figure 21: Output filter that is used to produce synthetic speech.

By artificially modifying the filter parameters, the NRC succesfully produced several voice modifications, including apparent sex changes, dialect changes, physiological changes (short larynx, large vocal tract, etc), and emotional changes, such as voices that convey fear, urgency, pain, or uncertainty. Other changes are obtained at the synthesis stage. For example, breathy voices are obtained by mixing together the two excitation sources (periodic signal and white noise), and whispered speech is obtained by using exclusively the white noise source.

CHAPTER 4

IMPROVEMENTS TO THE SEGMENTATION ALGORITHMS

4.1 Introduction

It may be worth restating here that the purpose of the research performed by the author was to improve the naturalness of the speech produced by the NRC vocoder. Although this system initially produced highly intelligible speech, researchers at the NRC suspected the naturalness of the speech could be improved. This chapter and the ones following it describe the research performed by the author to improve the naturalness of the speech.

The original system made several errors in estimating the fundamental frequency or misclassifying voiced frame as voiceless. Although segmentation decisions are not critical for the quality of the analysis (formant estimation), accurate pitch extraction and voicing decisions are necessary for a high quality synthesis. Recall that for the case of the pitch-synchronous analysis, the pitch-period is determined by the frame, and thus if a frame boundary is undetected, for example, the size of the current frame will be roughly twice of what it should be. The apparent pitch-period will instantly increase by 100%, and this will produce a "pop" in the synthesis. If several consecutive frame boundaries are undetected, the frames will be wrongfully classified as voiceless, producing even greater degradations.

Here is a list of the problems encountered with the original system:

- The program searched the time-differenced laryngograph signal for a positive peak —representing a vocal cord closure— above a preset (flat) threshold. If the program did not find a positive peak, it looked for a negative peak, usually corresponding the start of the open phase of the glottal cycle, using the same scheme. In the "h" sound in "three hundred," for example, the laryngograph signal for a given speaker shows only negative peaks (see Figure 22 p.52). If the program found a negative peak, it set the frame boundary at that position exactly, without taking into account the delay between an opening and a closing of the vocal cords. This resulted in an audible pitch variation.

- Within a voiced speech segment, the differentiated laryngograph signal often exhibits several spikes near instants of glottal closure. The original system often made the error of returning the location of these spikes instead of valid excitation points.

- The original system was also sensitive to isolated noise spikes on the laryngograph signal *i.e.*, noise spikes outside voiced regions. These spikes would trigger segmentation decisions if they were strong enough. If a frame in the middle of a silent region is declared as voiced because of a noise spike, the synthesis contains a "pop" sound at that location.

- Other problems occurred at the end of voiced regions, where vocal cord activity is often too weak to be detected by the laryngograph. This

automatically meant that the frame size was set to 10 ms. The speech power in this frame was then evaluated, and if it was above a threshold, the ratio of the first two autocorrelation coefficients was computed. If this ratio exceeded another threshold —indicating predominant low-frequency energy— the frame was classified as voiced, with a pitch-period of 10 ms. If the ratio was too low, the frame was classified as voiceless. If the power was too low, the frame was classified silent. Perhaps treating the frame as voiced with a pitch-period of 10 ms was better than treating the frame as voiceless, but this often causes voiced regions to end with a distinctive "click" sound.

- Sometimes, sudden movements of the larynx result in a large low-frequency component on the laryngograph signal, which in turn overloads the A/D converter in the PCM/video recording system. When this happens, the several frames are wrongly classified as voiceless. This problem could have been avoided by high-pass filtering the laryngograph signal, but some recordings were made to study the details of the laryngograph waveform, and filtering was not carried out because of the distortions it produces.

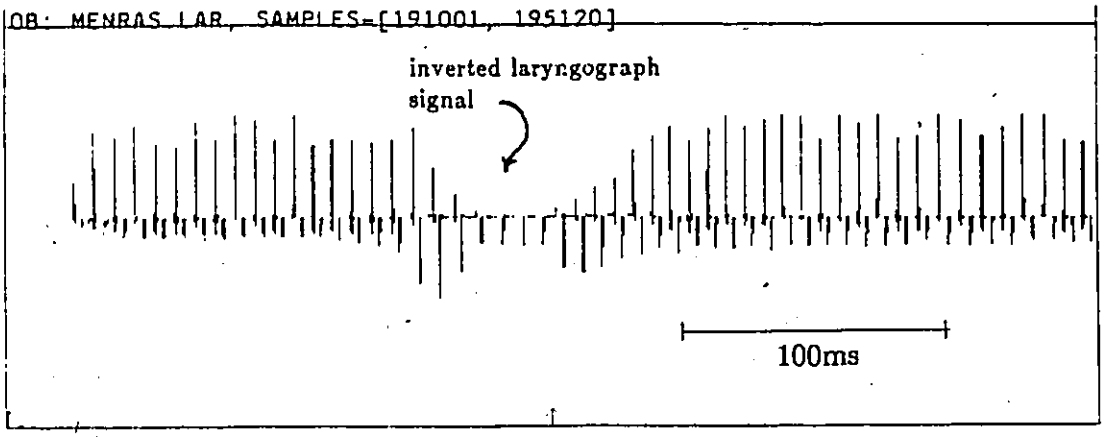

Figure 22: The differentiated laryngograph signal for the words "three hundred." The laryngograph signal is inverted in the "h" sound in "hundred."

To resolve all these problems, the author made two major modifications the original speech segmentation algorithm. He first changed the algorithm that extracts the pitch from the differentiated laryngograph signal. This includes the addition of a so-called U-shaped threshold to detect peaks in the differentiated laryngograph signal. He then added an autocorrelation pitch estimator at the end of voiced regions. Other modifications, which are considered to be minor, are discussed in Appendix A.

4.2 U-shaped threshold

The original pitch extraction algorithm used a flat threshold to detect spikes in the differentiated laryngograph signal. The improved pitch extraction algorithm uses two schemes to detect these spikes. The first scheme, which operates on the onset of voiced regions, is a "start-up" scheme because it makes the least

assumptions about shape of the laryngograph signal. It is similar to the original scheme in the sense that it looks at the differentiated laryngograph signal for values that are above a fixed threshold. However, the scheme was modified in order to ignore noise spikes caused by the recording system, where a noise spike is defined as a single spike in the middle of a silent or voiceless region. If a spike is found in such a region, the algorithm checks to see if there is a second spike up to 21 ms later (21 ms is an arbitrary upper limit on the length of a glottal cycle). If a second spike is not found, the first spike is disregarded. If a second positive spike is found, the first one is considered valid and the routine returns the location of the first spike, and the distance between the first and second spikes. This is used to adjust the length of the current (voiceless) frame in such a way that its length is equal to the first pitch-period.

Once the first scheme has detected four consecutive voiced frames for which the pitch does not vary by more than 10%, the second scheme takes over. Instead of using a flat threshold, this scheme applies a "U" shaped threshold at the region where the next excitation is likely to be, based on the previous pitch-period. This is a reasonable way of positioning the U-shaped threshold because the pitch-period varies slowly. The excitation point is determined the following way: if a single sample point is above this U-shaped threshold, it is assumed to be the excitation point. If several samples are above the threshold, the program computes the ratios of sample value over threshold value and chooses the largest. The U-shaped threshold is shown in Figure 23.

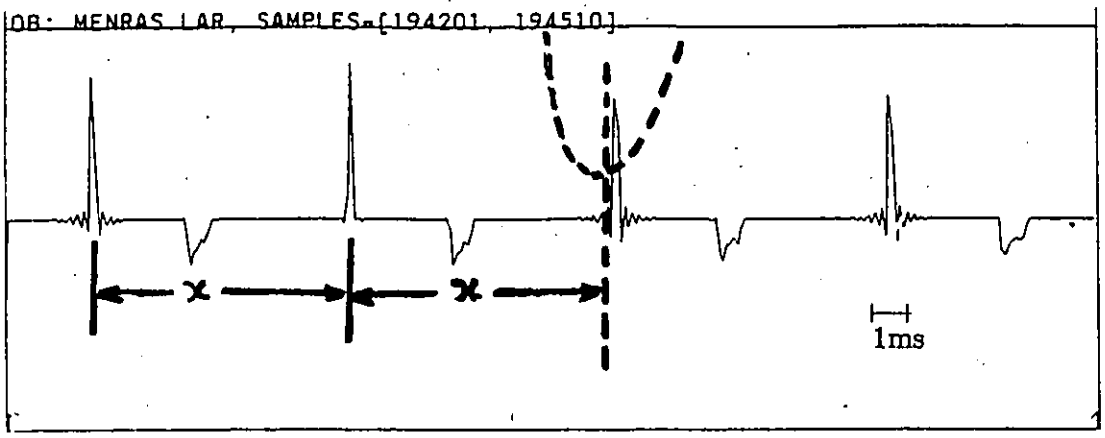

Figure 23: The U-shaped threshold placed on the differentiated laryngograph signal: the location of the centre of the U with respect to the beginning of the frame is equal to the previous pitch-period.

The U-shaped threshold was designed arbitrarily. It has its minimum value set at half the value of the flat threshold used by the first scheme, and rises with the cube of the distance from the center of the "U," with a doubling of the threshold value that occurs 15 samples away from the center of the "U."

4.3 Autocorrelation pitch estimator

If no samples are above the threshold, the algorithm does not look for negative spikes, but instead uses an autocorrelation based pitch estimator. This takes care of instances where the laryngograph cannot detect weak vocal cord activity, such as for a few frames at the end of voiced regions. This is shown in Figure 24.

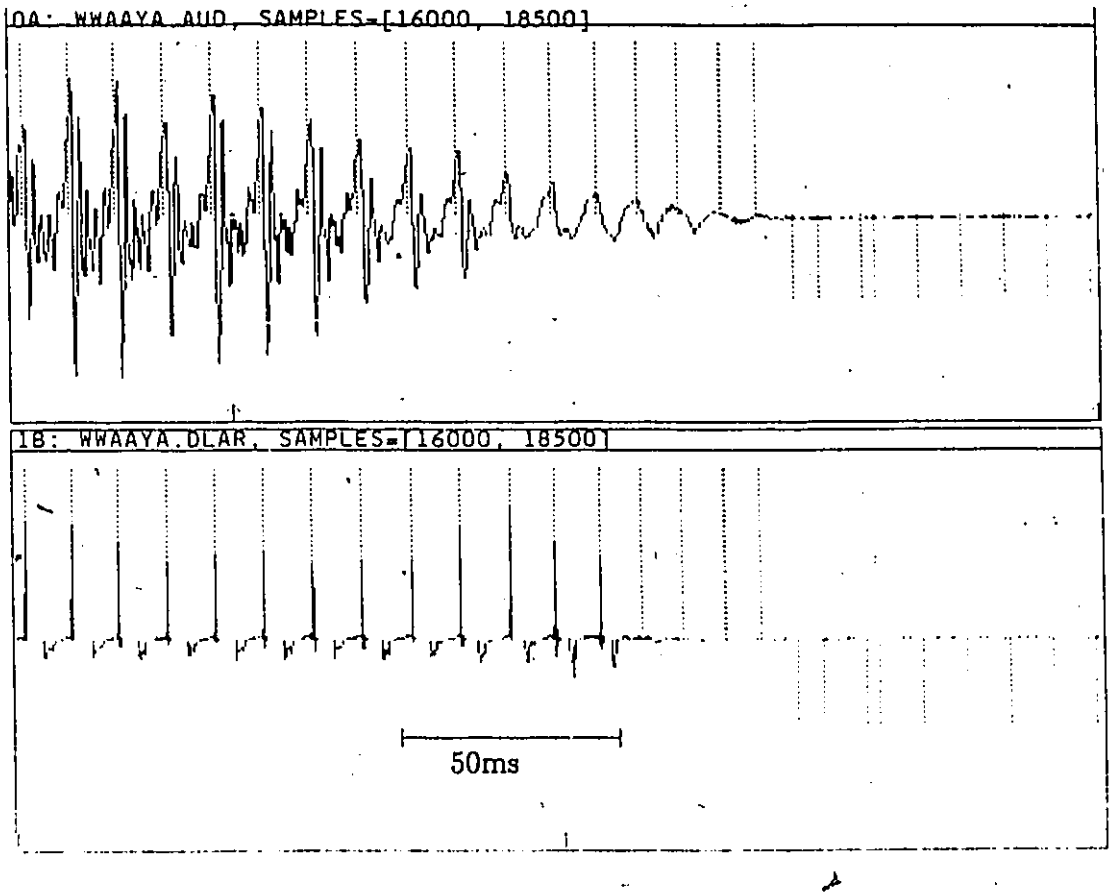

Figure 24: Autocorrelation pitch detection at the end of voiced regions: the top window shows the audio signal, and the bottom window shows the differentiated laryngograph signal. Both are time aligned and correspond the end of the spoken word "go" from the sentence "we were away a year ago." The tall vertical bars in both figures indicate the location of the frame boundaries as determined by the frame segmentation program: the bars going up to the top of the windows indicate voiced frames, and the bars going half way down indicate silent frames. As can be seen from these figures, the vocal cord activity at the end of voiced regions is often too weak to be detected by the laryngograph. The segmentation for the last four voiced frames was performed by the autocorrelation pitch extractor.

The autocorrelation pitch estimator also takes care of the cases where the laryngograph signal is clipped, as explained previously, or when the laryngograph

signal is inverted, such occurs as in certain voiced "h" sounds, as is sometimes the case in "hundred." The method computes autocorrelation coefficients on a low-pass filtered audio file, and the pitch is simply the index of the largest coefficient. Autocorrelation coefficients are computed as a sample mean:

$$r_j = E\{x(i)x(i+j)\} = \frac{1}{N}\sum_{i=1}^{N} x(i)x(i+j) .$$

A reasonable pitch found by this method is not a valid criteria to classify a frame as voiced. To discriminate voiced frames from voiceless and silent frames, the power in the filtered audio file is computed, and if the power is high enough (ruling out silent frames), the ratio of the first two autocorrelation coefficients is compared against a threshold. The autocorrelation ratio double checks the smoothness of the acoustic signal for this frame: a value close to "1" indicates that a predominance of low-frequency energy, and that the frame is likely to be voiced. A value close to "0" indicates that the frame is probably voiceless. A "reasonable" pitch-period means that the value of the current pitch-period should be close to that of the previous one, a condition tested by the inequality

$$.8 * oldp < per < 1.5 * oldp .$$

where *per* is the pitch-period found by the autocorrelation method and *oldp* is the previous pitch-period. This formula was defined arbitrarily, and takes into account the fact that the pitch can decrease faster that it can increase. The power threshold was heuristically determined, based on the background noise level on the tape recordings.

CHAPTER 5

EXCITATION SIGNALS FOR VOICED SPEECH

5.1 Introduction

LPC models the vocal tract as an all-pole filter, and because the output of a filter is a function of its input, the choice of excitation signal affects the quality of the synthesis. To produce voiceless speech, white noise is used as the excitation signal; this proves to be an accurate model of the air turbulence that causes such sounds, for the synthesis results are realistic. To produce voiced speech, the vocal tract filter is repetitively excited with a "driving function," and it is not evident which function will give optimal results. A "clean" impulse produces intelligible speech but causes a loss of the speaker's identity. The "Rosenberg" model of glottal air flow[28], a signal that approximates the air pressure gradient in the glottis, produces similar results, but with stronger low-frequency components. This signal is shown in Figure 25, and Figure 26 shows a doubly differentiated version.

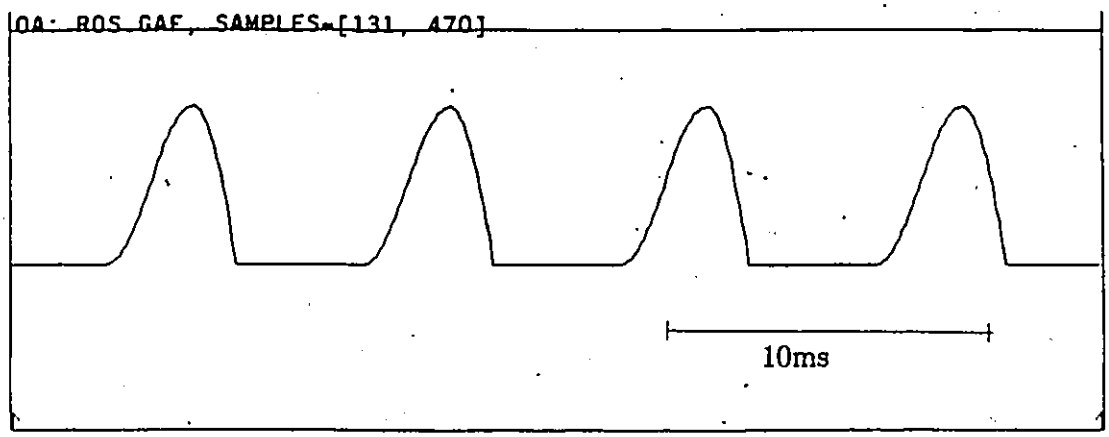

Figure 25: Rosenberg model of glottal air flow

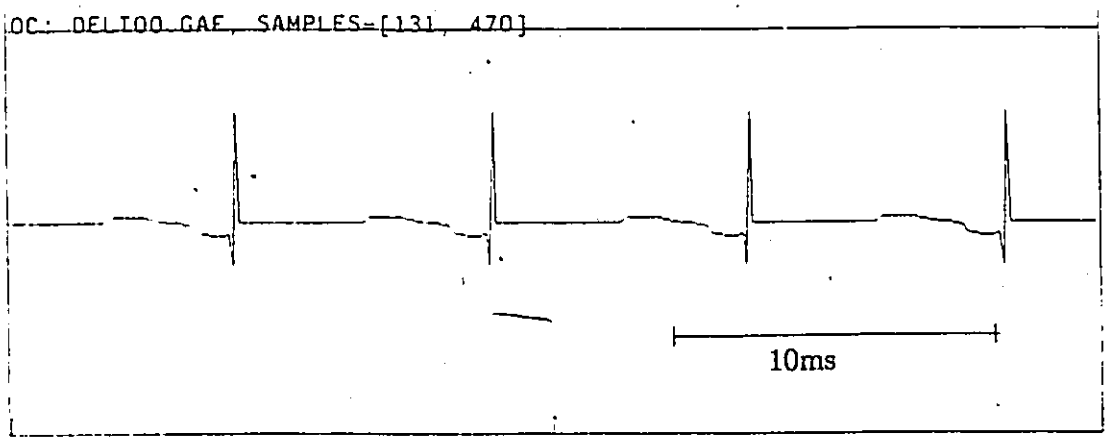

Figure 26: Doubly differentiated Rosenberg function

The original driving function in the LPC vocoder at the NRC was a residual from a single glottal cycle (arbitrarily chosen). Although this driving function produced speech of high quality, it was suspected that a more carefully chosen driving function would produce even better results. This Chapter

presents experiments that were carried out in an attempt to find such an excitation function.

5.2 Residual averaging

As mentioned in Chapter 1, speech synthesized from the entire residual signal is identical to the original audio signal: the reconstruction is exact. To find a good excitation function, a sensible thing to try is to average several voiced frames of a residual signal on a frame-to-frame basis. The hypothesis is that an averaging of residuals will cause irrelevant features to cancel out, and the mean residual will produce the best overall synthesis. Figure 27 shows a typical residual signal with the corresponding frame boundaries. Two methods are used to average residuals: time and frequency averaging.

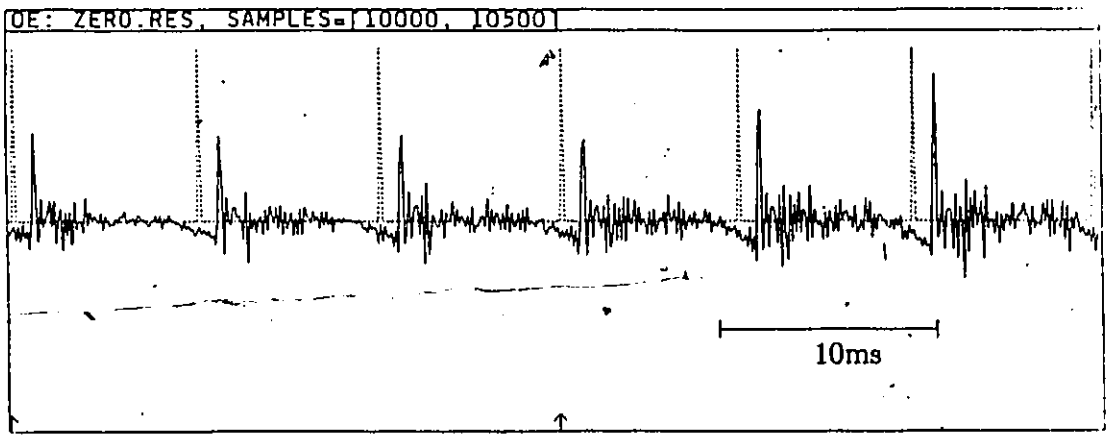

Figure 27: Typical residual signal from the LPC analysis of voiced speech . The dotted vertical lines indicate frame boundaries, which are synchronized with the glottal cycles. The residual signal is averaged on a frame-per-frame basis.

5.2.1 Time averaging of residuals: As seen from Figure 27, the residual contains major spikes near the beginning of a frame. Because the location of this spike varies from one frame to the next, residual frames must be "spike-aligned" before being time-averaged, since a direct frame-to-frame averaging will smear this spike. To be included in the average, it was decided that a residual frame must contain a major spike within the first 20 samples of a frame. This is an arbitrary number that depends on the delay between the speech signal and the laryngograph signal (the propagation delay of the acoustic signal from the vocal cords to the microphone).

Averaging many residuals this way was unsuccessful because the high frequency components jitter with respect to the impulse and cancel each other out in the averaging process to give a low-pass filtered excitation signal.

5.2.2 Frequency averaging: The jitter problem of the time averaging method can be avoided by applying a Fourier transform to each glottal cycle and averaging the amplitude and phase spectra separately before transforming back to the time domain.

The frequency averaging algorithm is more elaborate. First, the length of the longest frame is found; the remaining frames are then padded out with zeros to this length before the DFT's are computed for each frame. This causes all DFT's to have the same length, and eases the averaging process. Because phase is a relative quantity, averaging phase spectra only makes sense if the

phase is measured from a common reference point in all the frames. An obvious choice is the major spike at the instant of vocal cord closure (see Figure 27). The frequency averaging algorithm finds the location of this spike and passes it as a parameter to a special DFT program which adjusts the argument of the sine and cosine terms accordingly. (The "regular" DFT defines the phase relative to the first sample of the time signal).

Figure 28 shows the log amplitude spectra of averaged residuals for both methods. Unlike the time-averaging method, the spectral-averaging method is immune to phase jitter problems. However, the spectral averaging method is more susceptible to noise problems, since additive noise components on the residual will not tend to cancel out. The excitation obtained from this method was no better than using a residual from a single glottal cycle, and we have failed to improve on the original system in this respect.

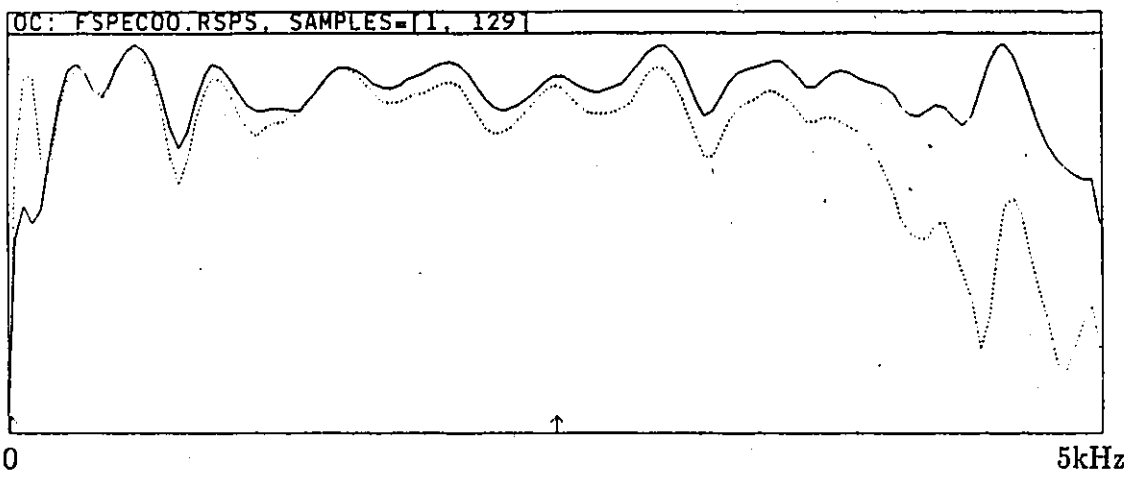

Figure 28: Log amplitude spectra of averaged residuals from the time averaging method (dotted line) and the frequency averaging method (solid line). For both of these plots, 150 frames were included in the average. Note the loss at high frequencies with the time averaging method.

5.3 Excitations derived from zero-phased signals

In the previous section, we took glottal residuals from a speaker, averaged them together, resynthesized the speech with this averaged residual, and determined if the resulting synthesis was improved.

This section describes an attempt to obtain a better driving function by setting to zero the phase spectrum of driving functions, while preserving their amplitude spectrum. This operation, which maximizes the "spikiness" of a signal, is desirable because a driving function with several major spikes will excites the vocal tract filter several times during a glottal cycle, producing a poor synthesis.

To zero the phase spectrum of a driving function, the DFT of the function is taken, and its amplitude and phase spectra are computed *i.e.*, the DFT coefficients are converted to polar form. All the angular coordinates are then set

to zero, and a polar-to-rectangular conversion is performed. Then, the IDFT of these rectangular coordinates is used to obtain the zero-phased time-signal.

Residuals from several voiced frames have been zero-phased this way. Figure 29 shows a typical residual from a voiced frame, and Figure 30 shows the zero-phased version of this signal.

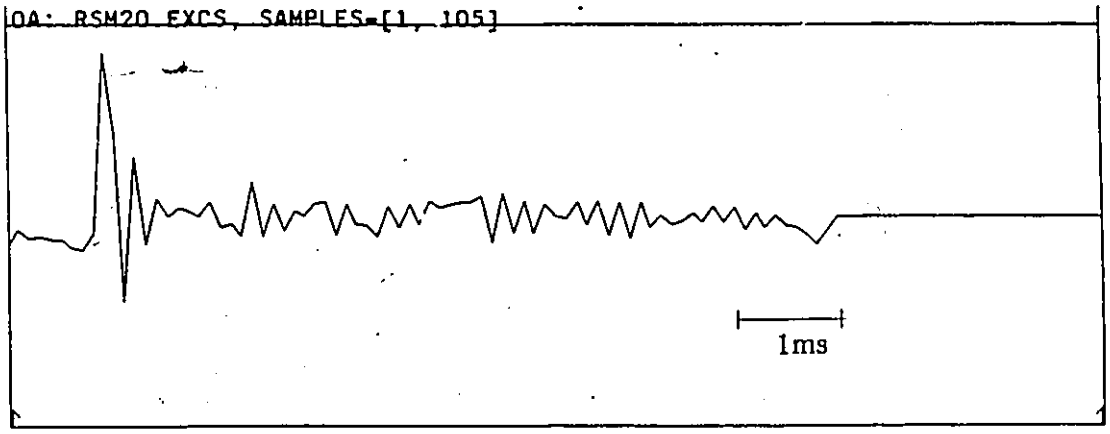

Figure 29: Residual of a voiced frame

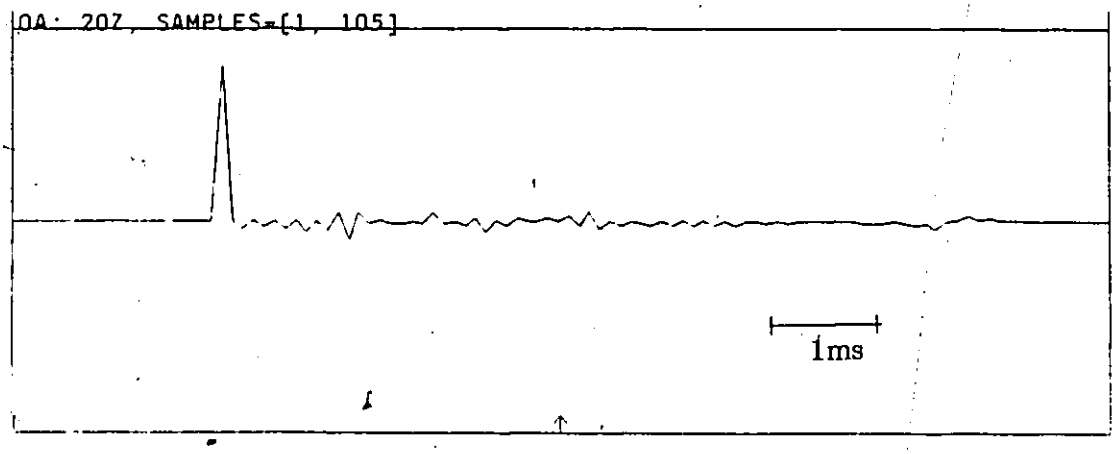

Figure 30: Same residual as in Figure 29, but with phase spectrum set to zero. Only the positive-time section (as shown in the figure) is used as the driving function,

Both signals look roughly the same, in the sense that major impulses are recognizable in both of them. The corresponding syntheses were similar, and no conclusions could be drawn.

Residuals from mixed-excited frames were also zero-phased in such a manner. With these residuals, zero-phasing must be used because, unlike voiced residuals, they usually do not have prominent spikes. The resulting syntheses contained only minor differences compared to the ones produced from voiced residuals.

A white noise signal was also zero-phased. Surprisingly, the syntheses it produced were similar to those produced from voiced residuals. From this we can conclude that a single frame-residual from a given speaker does not contain

"hidden" information necessary to obtain a superior synthesis from LPC parameters derived from his voice.

These experiments also showed the importance of the temporal structure (phase spectrum) of the excitations: periodically exciting the vocal tract with a noisy signal (a white noise sequence or a residual from a mixed-excited frame) produces a poor synthesis, but setting to zero the phase of both of these signals —while preserving their amplitude spectrum— produces a synthesis of comparable quality to that of a voiced residual.

From this section we can conclude that the amplitude spectrum of a good excitation must be generally flat and the temporal structure must contain a single major spike as well as a random component. To obtain such a signal, a reasonable approach is to choose a section of the residual signal that meets the above requirements. As previously mentioned, zero-phasing voiced residuals does change the tone of a synthesis in a minor way, but it is debatable whether naturalness is increased or not. The excitation function that is used at the NRC as a result of this work is a zero-phased residual from a voiced frame.

CHAPTER 6

SYNTHESIS OF VOICED FRICATIVES

6.1 Introduction

The most salient remaining problem with the synthesis was a "buzziness" in prolonged voiced fricatives. Voiced fricatives, or *mixed-excited* speech sounds, were defined in the first chapter as those whose source of acoustic energy is both the vocal cords and some turbulence noise from another constriction. Examples of voiced fricatives are the /z/ (alveolar voiced fricative), /v/ (labio-dental voiced fricative), and "zh" (palato-alveolar voiced fricative, such as in "beige"). With the original model, these sounds were labeled as voiced because the laryngograph detected vocal cord activity. As a result of this classification, the driving function used to synthesize these sounds was the one used for voiced speech, and the resulting synthesis contained a distracting buzziness. It was suspected that the problem could be resolved by exciting these sounds with a special excitation function, and this idea was tested by identifying the offensive portions manually and trying various excitation functions.

6.2 Choice of excitation function

The excitation signal used for voiced speech is a residual from a voiced frame, and hence, the appropriate excitation function for mixed-excited sounds was

thought to be a residual from a mixed-excited frame. Using such a residual did not improve the synthesis, but adding a white noise component to this excitation caused the buzziness to go away. It was subsequently discovered that the additive noise component must be random from cycle to cycle *i.e.*, cannot be a single sample function from a white noise process. We concluded that the impression of buzziness resulted not from the spectrum of the excitation but from its high correlation from glottal cycle to glottal cycle. In addition, further tests showed that a fixed percentage of white noise worked well for all voiced fricatives. (Some voiced fricatives are stronger than others; for example, the /z/ in *is* usually contain more frication than the /zh/ in *beige*.) More precisely, before adding white noise to the excitation signal, the excitation signal is scaled-down in order to have the ratio of white noise power over total power equal .8 .

Given that adding white noise to the excitation removes the perception of buzziness, one might question the need for using a residual from a mixed-excited frame as the basic driving function. There is, however, weak evidence that using a such a residual gives a more natural synthesis than using a residual from a voiced frame. As a result of this, a residual from a mixed-excited frame (from the /z/ in *is*) was later used to implement the synthesis of mixed-excited sounds.

6.3 Automatic detection of mixed-excited frames

To maintain the NRC's goal of a fully automatic analysis, we need an automatic

method to discriminate voiced fricatives from other voiced sounds. To this end, the following parameters were considered:

- the first autocorrelation coefficient of the audio signal. This parameter seemed appropriate, as it indicates a predominance of high frequency energy. The parameter was discarded, however, as it showed little correlation with mixed-excited sounds.

- the log amplitude spectra of both the audio signal and the residual signal. Of these two parameters, the log amplitude spectrum of the audio signal seemed more promising. There was little correlation between the high frequency contents of the residual and the presence of mixed-excited sounds.

- the average power in the signal obtained from subtracting two consecutive glottal cycle (frame) residuals. The hypothesis was that residuals from mixed-excited sounds change more from frame-to-frame than residuals from voiced frames. This measurement, however, gave no indication as to what type of region (voiced or mixed-excited) was being analyzed.

- the bandwidths of all formants.

- the residual power at the start of a glottal cycle (closed glottis interval) over the power for the whole glottal cycle.

The following three parameters were found particularly useful:

1. the total (broadband) power in the audio signal. More specifically, the parameter is the ratio of the power in a particular frame over the average power of all voiced frames of a file. During mixed-excitation sounds, this ratio is low.

2. the bandwidth of the first formant, B_1, estimated by the LPC analysis; it usually increases during mixed-excited sounds.

3. the ratio of energy above 3 kHz to total energy. This parameter, here termed the high frequency index (*hfi*), is usually high during mixed-excited frames due to the frication that occurs. A fourth order Butterworth filter was used to filter the audio signal to obtain a measure of the high-frequency energy.

The author established the *power* and B_1 parameters to be reliable indicators of voiced-fricatives by comparing LPC derived vocal tract parameters for both types of frames. Although the high frequency index was introduced by intuition, high frequency energy is commonly used to detect fricatives. However, the author found no references to the use of B_1 or of the broadband power to detect voiced fricatives.

The increase in B_1 was surprising, and it seemed possible that it might have been the response of the pitch-synchronous LPC analysis to the presence of noise in the speech. To test this hypothesis, we synthesized entire sentences with an excitation signal containing added noise and re-analyzed the sentence. The value of B_1 increased only slightly, less than the typical amount for voiced

fricatives. (see Figure 31). The increase of B_1 was therefore concluded to be a real property of the acoustic signal. Indeed, B_1 usually increases when the mouth is partially closed, such as during the phonation of voiced fricatives.

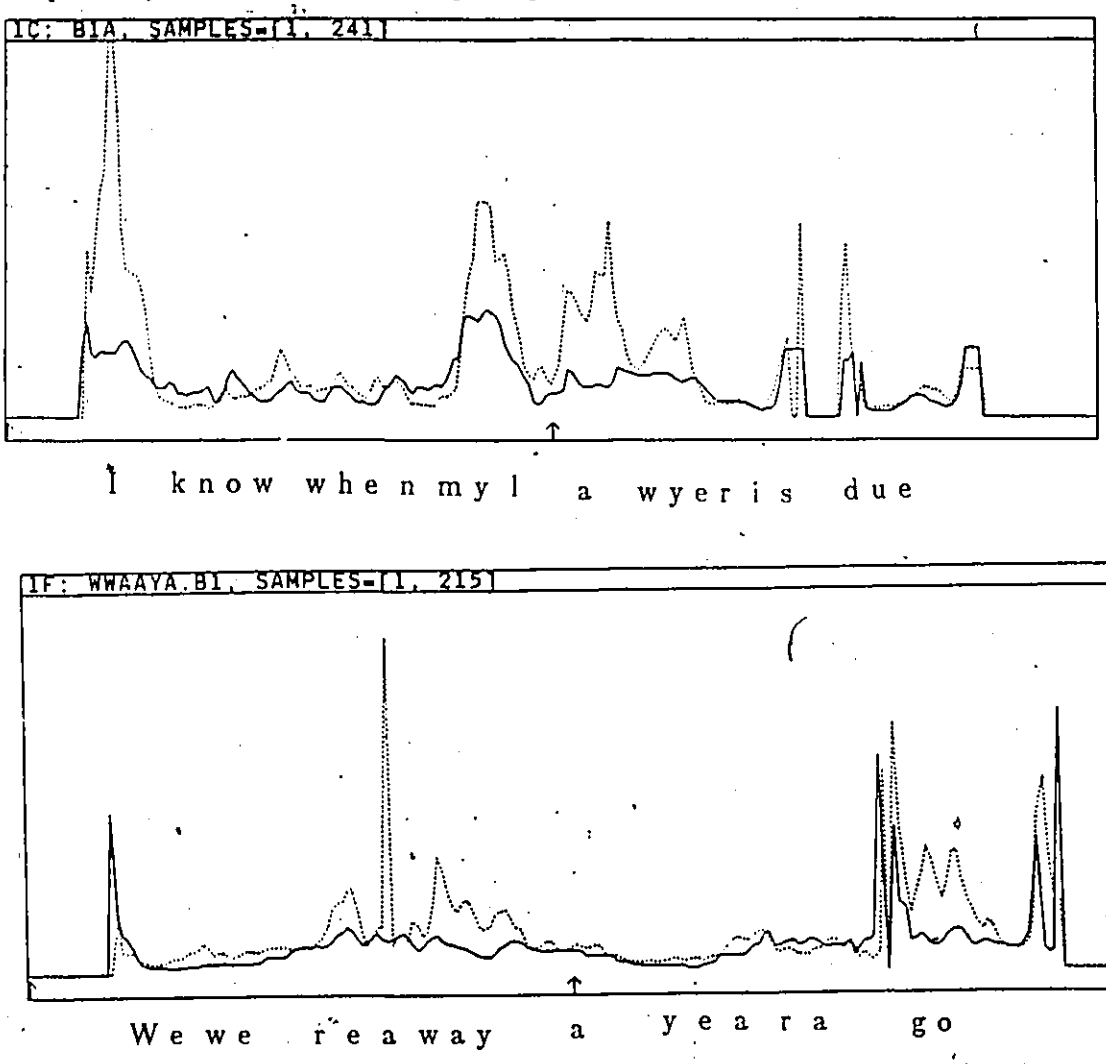

Figure 31: Effect of mixed-excitation on B_1. Both windows show the changing values of B_1 as a function of time. The top window is for *ikwmlid* and the bottom window is for *wwaaya*. In both windows, the solid lines are the original values of B_1 as obtained from a normal LPC analysis from the original digitized speech. The dotted lines were obtained by a) synthesizing the sentences with the mixed-excitation driving function, and b) re-analyzing this synthesis as if it were an original speech file.

Of the other formant bandwidths tested, such as B_2, B_3, and B_4, none

were found particularly useful. Although these bandwidths often increased in mixed-excited regions, this was not always the case, and hence did not provide a reliable indication of mixed excitation.

Figure 32 shows the acoustic waveform of the phrase "five seven" with the frame classification (vertical bars). Lines going to the top of the window indicate voiced frames, lines going half-way up, mixed-excited frames, lines going to the bottom, voiceless frames, and lines going half way down, silent frames. Figures 33, 34, 35 show the behaviour of all three parameters for this phrase.

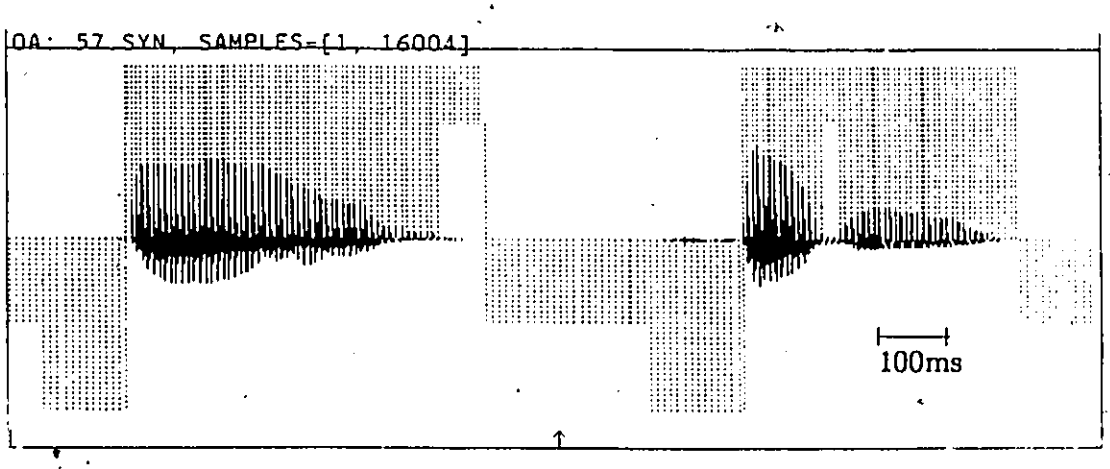

Figure 32: Acoustic waveform of "five seven" with framing information (vertical lines). Lines going to the top of the window indicate voiced frames, lines going half-way up, mixed-excited frames, lines going to the bottom, voiceless frames, and lines going half way down, silent frames.

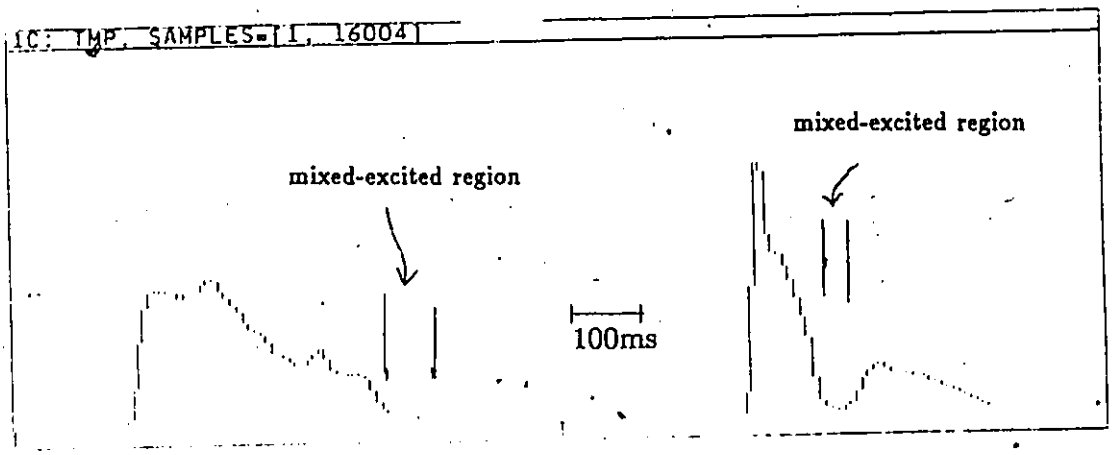

Figure 33: Power profile of "five seven."

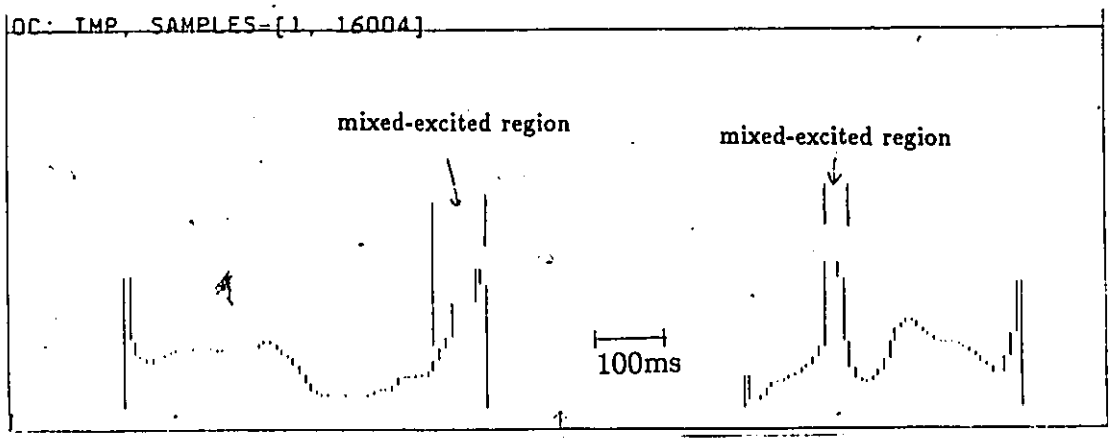

Figure 34: Bandwidth of the first formant for "five seven."

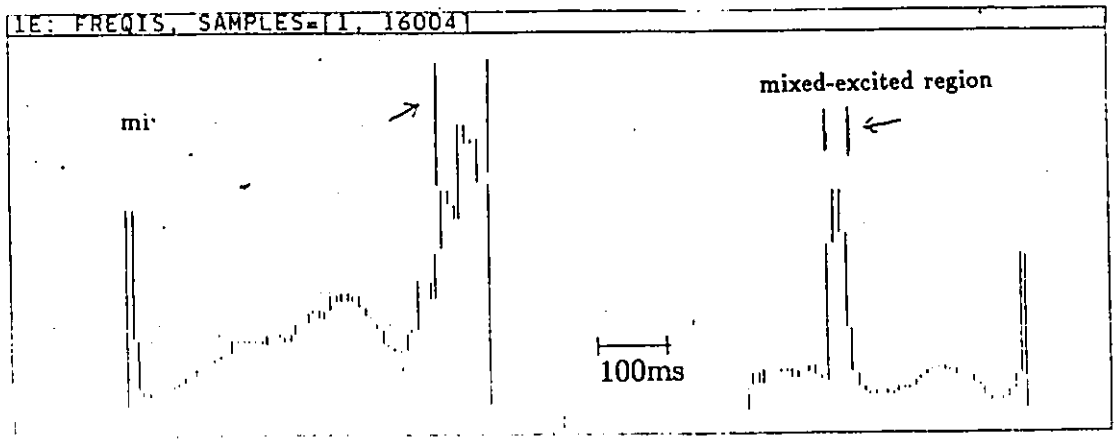

Figure 35: High frequency index of "five seven."

Since none of these parameters individually provided reliable discrimination, the author examined ways of combining them. The problem is to decide whether a frame is voiced or mixed-excited based on the value of these parameters. This is a binary hypothesis testing problem; the two hypotheses are that a frame is mixed-excited or it is not. Values from the set of parameters described in this section define a sample, and the optimum strategy, namely Bayes classification, is to assign a sample to the class which it is most likely to belong. This decision is based on estimates of class probability density functions, and on *a priori* class probabilities. The two hypotheses (or classes) are:

$$H_1 : \textit{the frame is a voiced fricative}$$
$$H_0 : \textit{the frame is not a voiced fricative}$$

Information on class probability distributions was obtained by manually identifying the voiced and mixed-excited frames contained in the following sentences, and noting the values of the *power*, B_1, and *hfi* parameters for both classes.

Speaker A:

1. "zero"
2. "five seven"
3. "I know when my lawyer is due."
4. "heading: left, one five two."

5. "zero one two three four five six seven eight nine ten"

6. "May we all learn the yellow lion roar."

7. "We were away a year ago."

8. "Altitude: fifteen thousand, two hundred and fifty."

9. "This is full of voiceless fricatives."

Speaker B:

1. "I know when my lawyer is due."

2. "We were away a year ago."

Speaker C:

1. "The beige Buick chugged along."

2. "Oops ! Pardon me Jack."

6.3.1 Quadratic classifier

A first attempt consisted of making parametric assumptions about class probability distributions: they were assumed to be multivariate Gaussian. Probability densities corresponding to the parameter values of a sample were estimated by computing centroids and covariance matrices for each class, and samples were classified assuming equal *a priori* class probabilities. Not surprisingly, this method proved unreliable, since individual parameter distributions were far from Gaussian (see Figures 40, 41 42 on pages 79, 80, 81), and class covariance

matrices and class centroids were found to be inconsistent from speaker to speaker. These matrices are shown in Figure 36, where *pow*, B_1, and *hfi* denote the three parameters being considered for the classification (power, bandwidth of the first formant, and high frequency index). Here, "unreliable" means that frame misclassifications produced perceptually salient errors. Empirically adjusting the *a priori* class probabilities failed to improve the performance of this classifier.

speaker A (voiced, correlation matrix)

pow	1.0000	-0.2780	-0.2303
b1	-0.2780	1.0000	0.3968
hfi	-0.2303	0.3968	1.0000

speaker A (voiced, mean vector)

1.03	78.86	0.10

speaker A (mixed-excited, correlation matrix)

pow	1.0000	-0.4140	-0.4671
b1	-0.4140	1.0000	0.1509
hfi	-0.4671	0.1509	1.0000

speaker A (mixed-excited, mean vector)

0.06	187.64	0.44

pow	b1	hfi

speaker B (voiced, correlation matrix)

pow	1.0000	-0.2854	-0.0957
b1	-0.2854	1.0000	-0.1421
hfi	-0.0957	-0.1421	1.0000

speaker B (voiced, mean vector)

1.05	121.21	0.11

speaker B (mixed-excited, correlation matrix)

pow	1.0000	0.1832	-0.0915
b1	0.1832	1.0000	0.3700
hfi	-0.0915	0.3700	1.0000

speaker B (mixed-excited, mean vector)

0.25	190.80	0.31

pow	b1	hfi

Figure 36: Correlation matrices and mean vectors for two speakers. Voiced data is on the top, and mixed-excitation data is on the bottom. For a given matrix, from left to right and top to bottom, the numbers correspond to parameters *pow*, B_1, and *hfi*. Also shown are the mean vectors for both types of frames (voiced and mixed-excited).

6.3.2 Perceptron Classifier

The next attempt consisted of training a multi-layer perceptron[29] on the discrimination task. Perceptrons are a class of neural-net classifiers, and neural nets consist of computational elements, called nodes, which are intercon-

nected via weights, mimicking the behavior of biological neural nets. The output of a given node is a weighted sum of its inputs (outputs from previous nodes) passed through a nonlinearity (see Figure 37). Generally, the weights are adapted during use to improve performance.

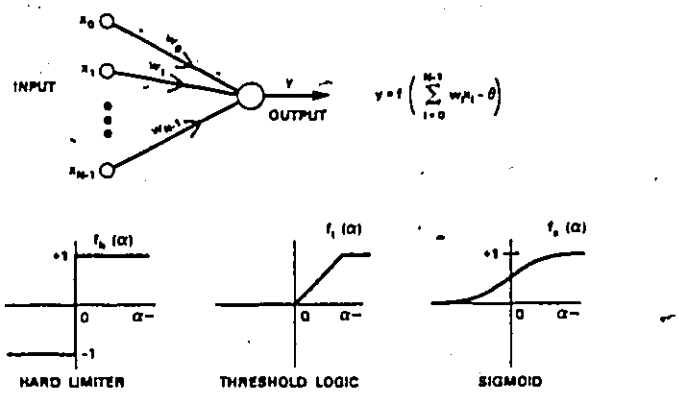

Figure 37: Computational element (from Lippmann[30])

A perceptron classifier is an artificial neural net whose input values are continuous (analogue), and whose training is done under supervision. Multilayer perceptrons contain several layers of nodes: an input layer, an output layer, and a variable number of intermediate (hidden) layers. Training data is presented to the input and output nodes, and the node weights are typically adjusted according to the backward error propagation algorithm. This algorithm makes no assumptions about the underlying distribution of the data but concentrates on minimizing the error when the distributions overlap. To train perceptrons, corresponding data pairs are presented at the input and output nodes, and the weights are adjusted according to the back-propagation algorithm. Once the training is over, the

values of the weights are fixed.

Different neural net topologies affect the outcome of the classifications. The number of layers determines the way in which the observation space can be divided. A single layer forms a half-plane decision region. A perceptron containing two layers can form convex decision regions, where convex means that a straight line connecting two points on the border of a region lies entirely within that region. A three-layer perceptron can form arbitrarily complex decision regions. The number of nodes in the hidden layers determines the complexity of the decision regions that can be formed within the classes just mentioned. Lippmann[30] states that no more than three layers are required with this type of feed-forward net because a decision region of any complexity can be generated if the number of nodes within the layers are sufficient. The number of nodes must be high enough as required by the input data, but if their number is too large, the weights may not be properly estimated from the training data.

The perceptron program used for the voiced fricative discrimination was written by Melvyn Hunt, my thesis supervisor, and Claude Lefèbvre, a colleague at the NRC. This program uses the Back-Propagation training algorithm to adjust the weights. The non-linear function used is the sigmoid, which is a function of the type

$$f(x) = \frac{1}{1+e^{-kx}} \qquad (18)$$

The sigmoid function is shown in Figure 38, with k=10, and x $\in$ [-1, 1]. The sigmoid function controls the dynamic range of the output of individual nodes.

As well, it introduces a non-linearity in the system. A multi-layer perceptron that employs a non-linear weight function can be viewed as a multi-stage non-linear adaptive filter.

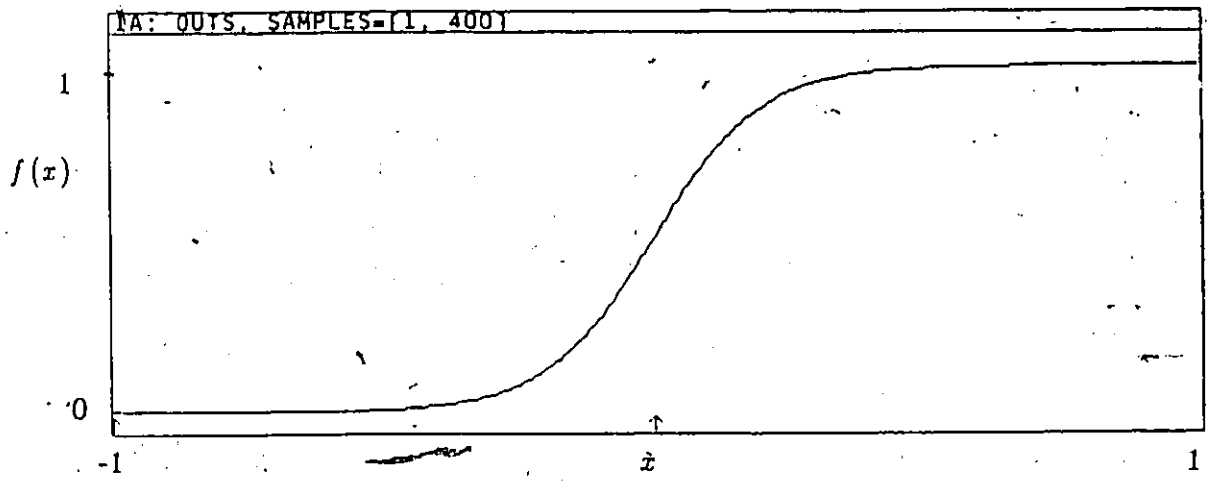

Figure 38: Sigmoid function

The input training data consisted of sets of parameter values from both voiced and mixed-excited frames; common scaling factors were used to bound the values between "1.0" and "-1.0" The output training data were either "1" or "-1," depending on whether the corresponding input samples were from voiced or mixed-excited frames respectively. The Back-Propagation algorithm, which is iterative, was left to run until the values of the weight stabilized (~ 800 iterations) for the topologies shown in Figure 39.

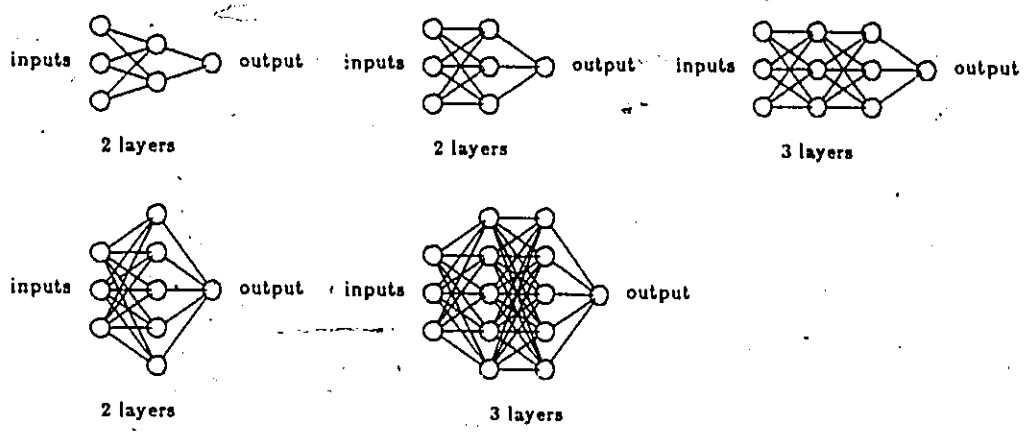

Figure 39: Perceptron topologies tested

All topologies were trained using identical test and training data, and all performances were similar to that of the quadratic classifier: there were instances where voiced glottal cycles without frication were erroneously judged to contain frication, and *vice versa*. Because these errors were noticeable in the re-synthesis, no other testing was performed, and the perceptron classifier was disqualified.

6.3.3 Statistical Classifier

The last attempt was more empirical. Histograms (Figures 40, 41, and 42) of the parameter distributions were fitted with simple functions, or in some cases with a sequence of straight lines. A note regarding these histograms:

because there are fewer mixed-excited frames than there are voiced frames, given a set of digitized sentences, there will always be more information on the distribution of voiced frames than on that of mixed-excited frames: mixed-excited histograms were computed from 150 frames, whereas voiced histograms were computed from 2000 frames.

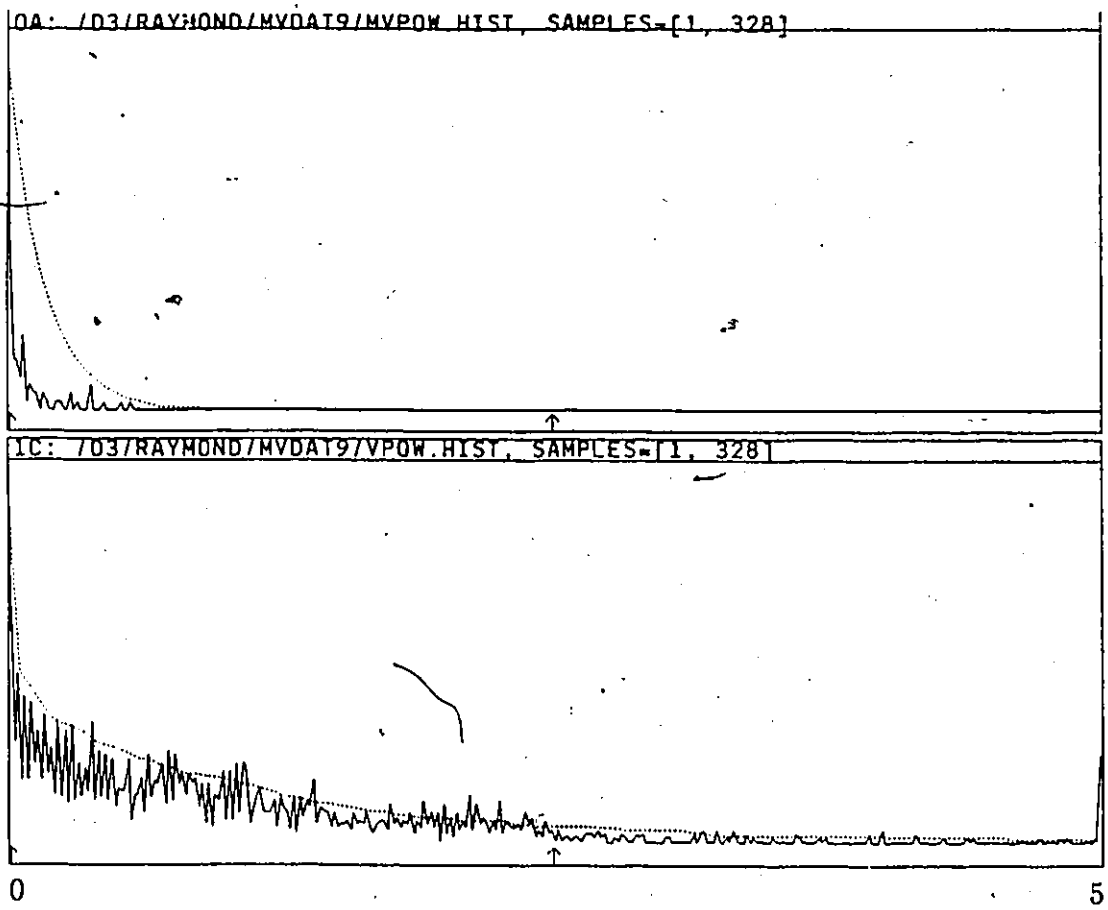

Figure 40: Power histograms: mixed-excited (top), voiced (bottom). The solid lines are the histograms, and the dotted lines are the shape of the corresponding density functions used by the statistical classifier.

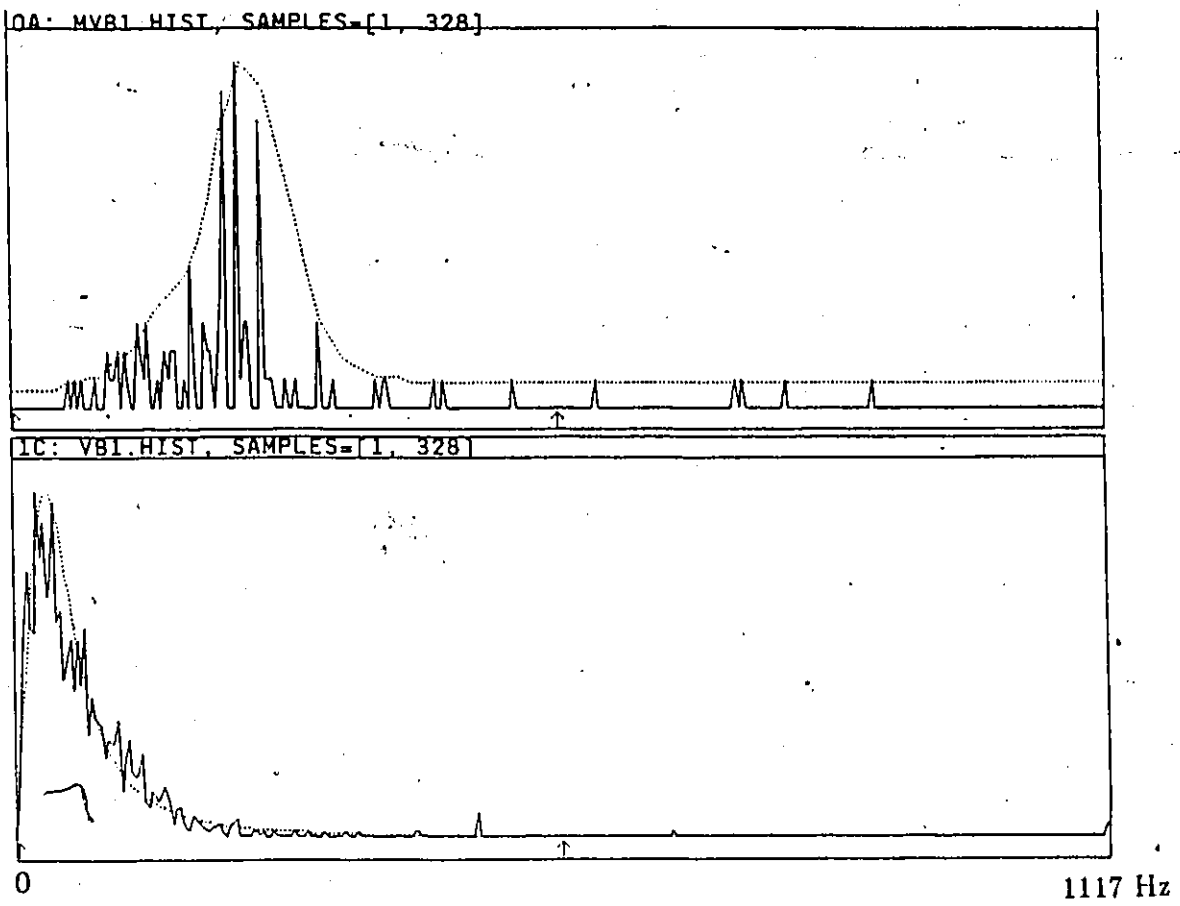

Figure 41: B_1 histograms: mixed-excited (top), voiced (bottom). The solid lines are the histograms, and the dotted lines are the shape of the corresponding density functions used in the statistical classifier.

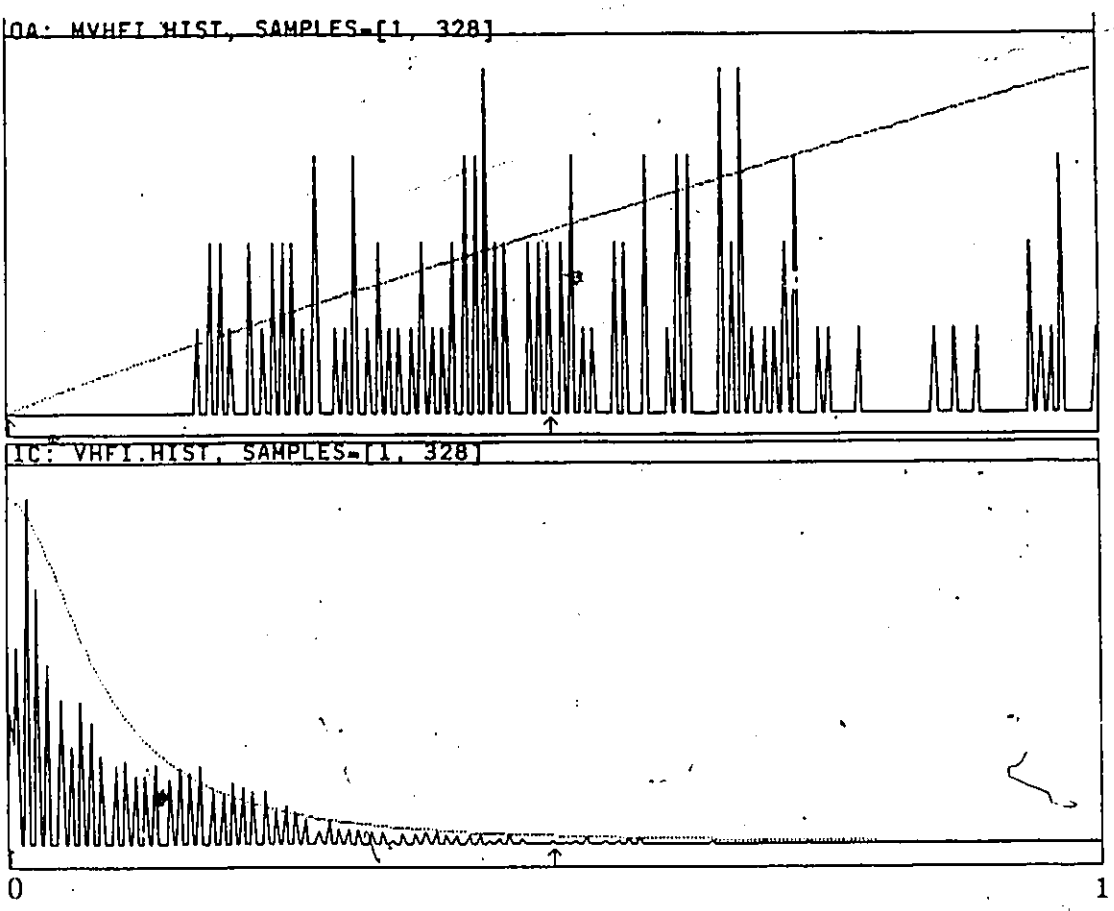

Figure 42: high frequency index histograms: mixed-excited (top), voiced (bottom). The solid lines are the histograms, and the dotted lines are the shape of the corresponding density functions used in the statistical classifier.

Once the proper shapes were obtained, functions and tables were scaled in order to equate their areas to unity. The scaled functions were as follows:

$$f(pow|H_1) = 6e^{-6x}, \quad x \in [0,\infty]$$
$$f(pow|H_0) = \textit{a look-up table}$$

$$f(B_1|H_1) = \textit{a look-up table}$$
$$f(B_1|H_0) = 1.735531 \times 10^{-6} \frac{319.4888}{1+319.4888x^{3.5}}, \quad x \in [0,\infty]$$

$$P(hfi|H_1) = 1.9x^{0.9}, \quad x \in [0,1]$$
$$P(hfi|H_0) = \frac{7.797}{1+150x^2}, \quad x \in [0,\infty]$$

Figures 43, 44, and 45 show the linear and logarithmic probability density functions (pdf's) of all three parameters. The solid lines are mixed-excitation pdf's, and the dotted lines are voiced pdf's.

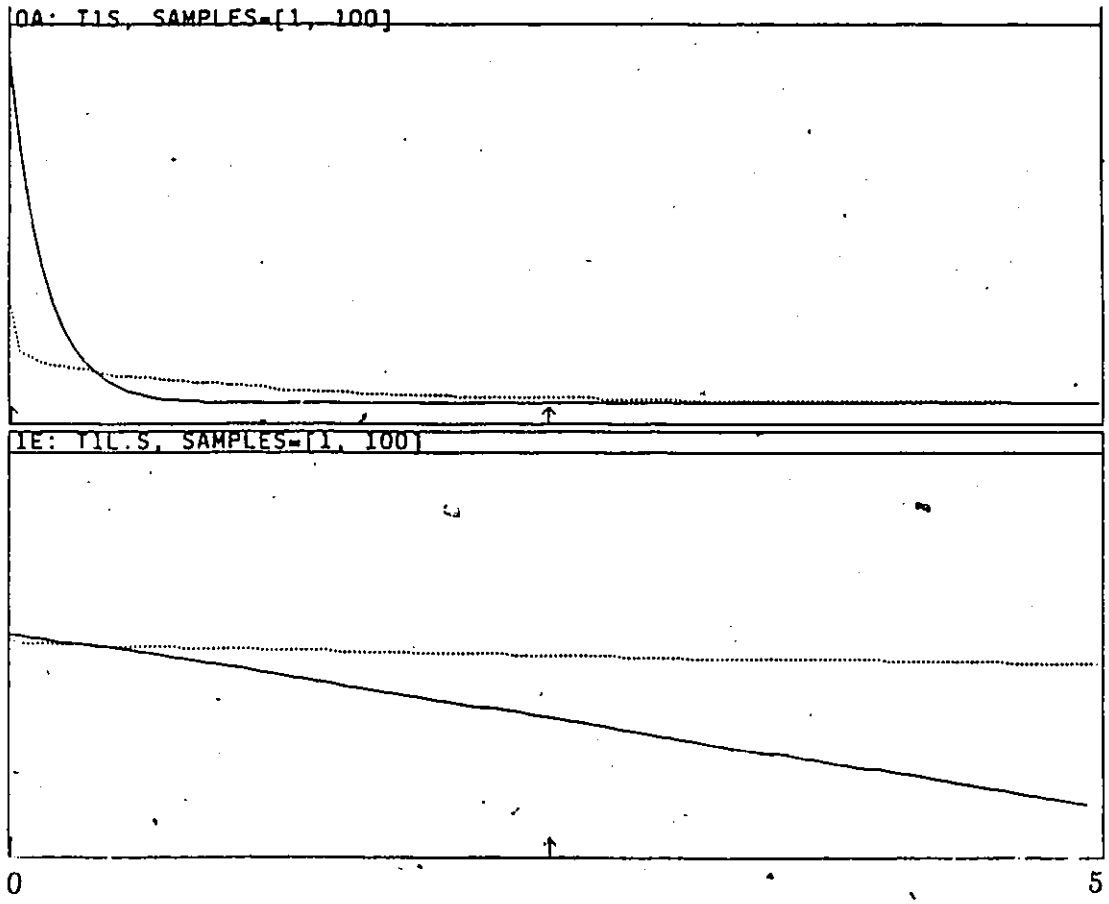

Figure 43: Linear (top) and log (bottom) pdf's of the *pow* parameter. The solid lines are mixed-excitation pdf's, and the dotted lines are voiced pdf's

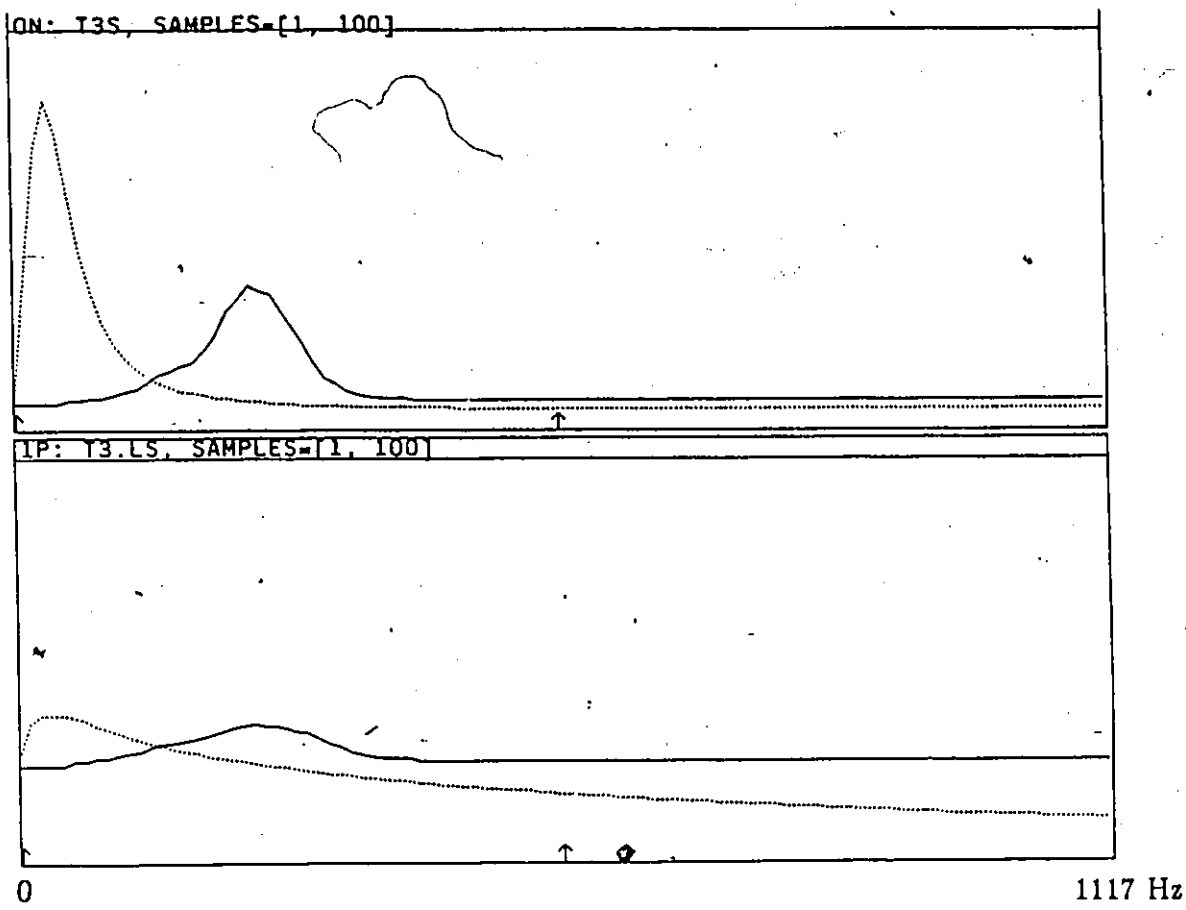

Figure 44: Linear (top) and log (bottom) pdf's of B_1. The solid lines are mixed-excitation pdf's, and the dotted lines are voiced pdf's.

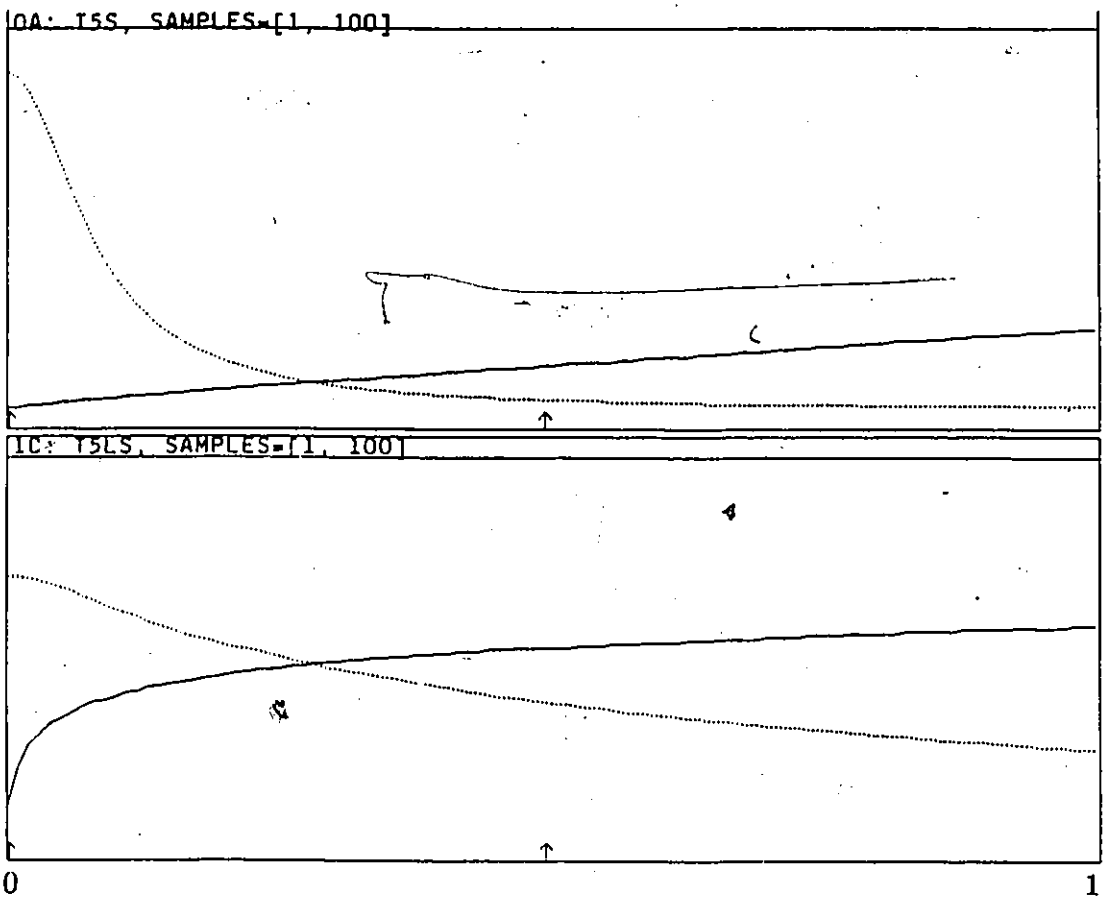

Figure 45: Linear (top) and log (bottom) pdf's of the hfi parameter. The solid lines are mixed-excitation pdf's, and the dotted lines are voiced pdf's.

Although a concern for determining these density functions was to fit the histograms as closely as possible, another one was the behavior of these functions in regions where data was sparse. For example, the density function $f(pow|H_1)$ in Figure 38 does not seem to fit the histogram well; here, closeness of fit was sacrificed for the simplicity and the rate of decrease of the chosen exponential function. The rate of decrease is important because as a parameter value tends towards and an extreme, its density functions should increasingly favour one

hypothesis over the other.

Due to lack of contrary evidence, all three parameters are assumed statistically independent. This is a reasonable assumption because the off-diagonal terms of the correlation matrices are not strikingly similar from speaker to speaker (see Figure 36). Prior to classifying a sample, *a priori* class probabilities are required ($Pr(H_0)$ and $Pr(H_1)$). Since most sentences contain more voiced frames than mixed-excited frames, a logical decision would be to assign a higher probability to H_0. However, as a starting point, both probabilities were assumed equal (.5). The choice of these values only affect the decision threshold, and this threshold can empirically adjusted if necessary.

Given these assumption, class probabilities are estimated according to Bayes' theorem:

$$Pr(H_1|\text{pow}=pow,\ \text{B}_1=\text{B}_1,\ \text{hfi}=\text{hfi}) = \frac{\text{f}(\text{pow}|\text{H}_1)}{\text{f}(\text{pow})}\frac{\text{f}(\text{B}_1|\text{H}_1)}{\text{f}(\text{B}_1)}\frac{\text{f}(\text{hfi}|\text{H}_1)}{\text{f}(\text{hfi})}\text{Pr}(\text{H}_1) \qquad (19)$$

$$Pr(H_0|\text{pow}=pow,\ \text{B}_1=\text{B}_1,\ \text{hfi}=\text{hfi}) = \frac{\text{f}(\text{pow}|\text{H}_0)}{\text{f}(\text{pow})}\frac{\text{f}(\text{B}_1|\text{H}_0)}{\text{f}(\text{B}_1)}\frac{\text{f}(\text{hfi}|\text{H}_0)}{\text{f}(\text{hfi})}\text{Pr}(\text{H}_0) \qquad (20)$$

where $f(x)$ denotes the value of a probability density *i.e.*, $f(x) = f_{\text{x}}(\text{x}=\text{x})$, with $f_{\text{x}}(\text{x})$ denoting the probability density function of the random variable x. $f(x|H_i)$ is a density value assuming H_i is valid, and $Pr(H_i|\text{x}=\text{x})$ denotes the probability of hypothesis i being valid based on the observed sample value x. The test then is to choose H_1 if

$$f(pow|H_1)f(B_1|H_1)f(hfi|H_1)Pr(H_1) \geq f(pow|H_0)f(B_1|H_0)f(hfi|H_0)Pr(H_0)$$

or

$$\Big(\ln[f(pow|H_1)]+\ln[f(B_1|H_1)]+\ln[f(hfi|H_1)]\Big) - \Big(\ln[f(pow|H_0)]+\ln[f(B_1|H_0)]+\ln[f(hfi|H_0)]\Big) \geq \ln[Pr(H_0)]-\ln[Pr(H_1)]$$

Figure 46 compares the performance of this classifier against that of the multi-layer perceptron. Shown are "decision curves," which are the output of both types of classifiers *i.e.*, the value of

$$\ln\Big[Pr(H_1)\Big] - \ln\Big[Pr(H_0)\Big]$$

for all voiced frames. If a decision curve is greater than zero (indicated by the horizontal axis), the frame is classified as mixed-excited. The top window contains decision curves for the phrase "*five seven,*" and the bottom window shows decision curves for the sentence "*the beige Buick chugged along.*" In both windows, the solid line is the decision curve from the perceptron classifier, and the dotted line is the output from the statistical classifier. The perceptron classifier, which had a single hidden layer containing two nodes, was trained with the back propagation algorithm using 700 iterations. For either type of classifier, the test and the training data were identical.

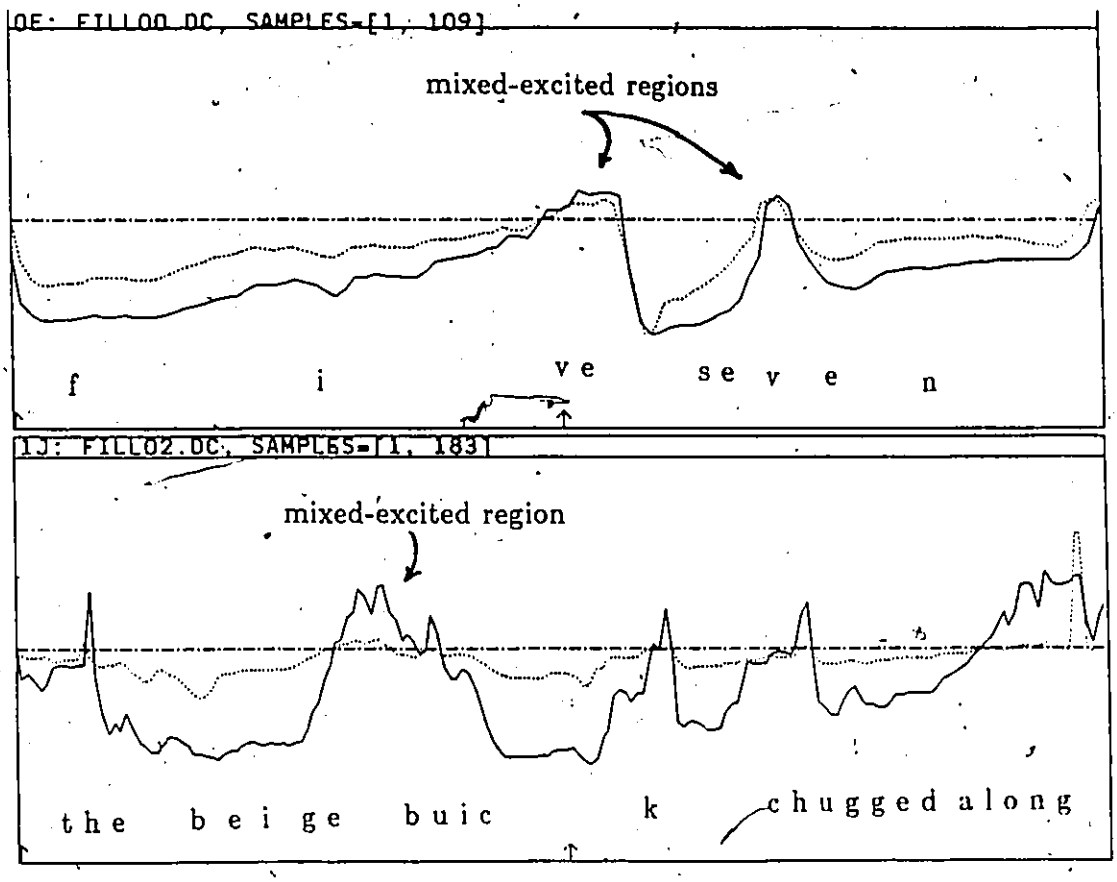

Figure 46: Typical decision curves from the statistical (dotted lines) and perceptron classifiers. If a decision curve is greater than zero (indicated by the horizontal axis), the frame is classified as mixed-excited. The top window contains decision curves for the phrase "*five seven,*" and the bottom window shows decision curves for "*the beige Buick chugged along.*" For both types of classifiers, these test sentences were included in the training data.

As with the multi-layer perceptron classifier, the statistical classifier also made salient errors. Although empirical adjustments of the *a priori* class probabilities were unsuccessful, it was possible to eliminate these errors by submitting glottal cycles judged to contain frication to an additional test using only the relative energy above 3 kHz (the so-called high frequency index parameter). Although this

filters out salient errors, it also removes a few glottal cycles that had previously been identified correctly as mixed-excited. However, since they did not have much high frequency energy, they do not sound buzzy when excited with the standard voiced excitation signal. To recapitulate, then, it was finally decided to judge a frame as belonging to a voiced fricative if $Pr(H_1) \geq Pr(H_0)$, and if

$$\ln\left[P(hfi|H_1)\right] - \ln\left[P(hfi|H_0)\right] \geq 0$$

computed with *a priori* class probabilities set at 0,5 .

6.3.4 Speech data for the evaluation of the statistical classifier

Because it did not make salient errors with the training data (the sentences listed at the beginning of section 6.3), this method was tested on a new set of sentences spoken by a new speaker (sentences that were not included in the training data). These sentences were chosen to contain a high proportion of voiced fricatives and included examples of all the voiced fricatives found in English:

1. "Five and seven is twelve."
2. "Eve drove the car into the edge of the grove."
3. "Those people bathe for pleasure in their leisure time."
4. "At the edge of the hedge lives a large beige dove."

5. "This judge with the badge on his breast always grudges against the church."

6.3.5 Results

No major errors could be detected either by listening or by examining the locations of the regions declared to contain voiced frication. A careful observation of classification decisions revealed that minor errors did occur, but these errors were not perceptually detectable for they occurred mostly in regions of low acoustic power. As well, "borderline" frames can usually be classified as voiced or mixed-excited without adverse effects.

6.4 Implementation

Figure 47 shows how the NRC's speech production model was modified for the synthesis voiced fricatives. It now contains two periodic excitation signals, signal "a" for voiced sounds, and signal "b," for voiced fricatives. Signal "a" is a residual from a voiced frame, and signal "b" is a residual from a mixed-excited frame (one which also has a major spike). A fixed amount of white noise is added to signal "b"; this was found to be crucial to the synthesis voiced fricatives.

The choice of excitation is encoded separately, but not independently, from the output file of the LPC analysis: for each speech frame, the synthesis

program reads a single parameter from a so-called mixed-voicing file. In this file, a "1" indicates a mixed-excited frame *i.e.*, signal "b" should be used along with the additive noise component. A "0" indicates a purely voiced frame (periodic signal "a" is to be used).

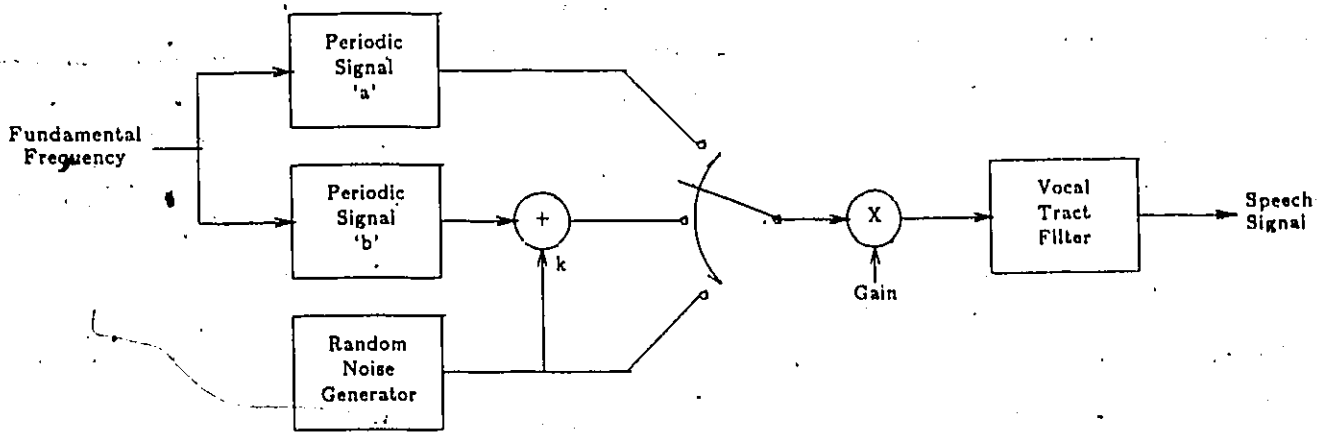

Figure 47: Speech production model for voiced, voiceless, and mixed-excited speech sounds.

CHAPTER 7

PITCH-ASYNCHRONOUS LPC

During work done for this thesis, it was observed that for a pitch-synchronous analysis using the covariance method, the quality of the analysis is not affected by offsets of the analysis interval with respect to the glottal cycle. This is surprising because this property had been considered valid only for much longer analysis intervals that include several pitch-periods[23]. The finding led to the observation that with the covariance method, pitch-asynchronous analyses using 10 ms windows are comparable to pitch-synchronous analyses. This chapter is an inquiry into the mechanisms behind these observations.

For voiced frames, LPC analyses carried out with short analysis intervals (intervals at least 25% shorter than the pitch-period) are sensitive to the placement of the analysis window the within the glottal cycle. One reason for this is that during the open glottis phase, the trachea and lungs are acoustically coupled to the vocal tract, causing a lowering of formant frequencies and an increase in formant bandwidths. Hence, the acoustic properties of the vocal tract vary during intervals of glottal cycles.

Because of this, pitch-asynchronous analyses are generally carried out with long windows, and are therefore considered less precise, since the LPC prediction error is minimized over the entire analysis interval. Although continuity of

formant tracks have been found insensitive the placements of such windows within glottal cycles[31], continuity of formant tracks is not a valid criterion for the accuracy of the analysis, and typical problems with pitch-asynchronous analyses in general include poor bandwidth estimates and a shift in the first formant frequency caused by harmonics of the fundamental. Another difficulty of pitch asynchronous analyses perhaps worth mentioning is the estimation of the excitation power, since excitation instants will not generally fall at the start of an analysis intervals, and a given interval may contain several excitation points. However, a "bug" in the original system caused significant errors in power estimates, and correction of the problem did not improve the perceived quality of the synthesis. This leads us to conclude that, to some extent, we are perceptually insensitive to absolute power levels in speech.

During the research performed for this thesis, it has been observed that pitch-asynchronous analyses using short (10 ms) windows produces formant frequency estimates comparable to those from pitch-synchronous analyses, and provided that the bandwidth of the first formant (B_1) is not allowed to decrease below a threshold, the resulting synthesis is comparable if not better than a synthesis obtained from a pitch-synchronous analysis. Figure 48 shows the closeness between formant frequencies tracks derived from the pitch-synchronous and pitch-asynchronous analyses.

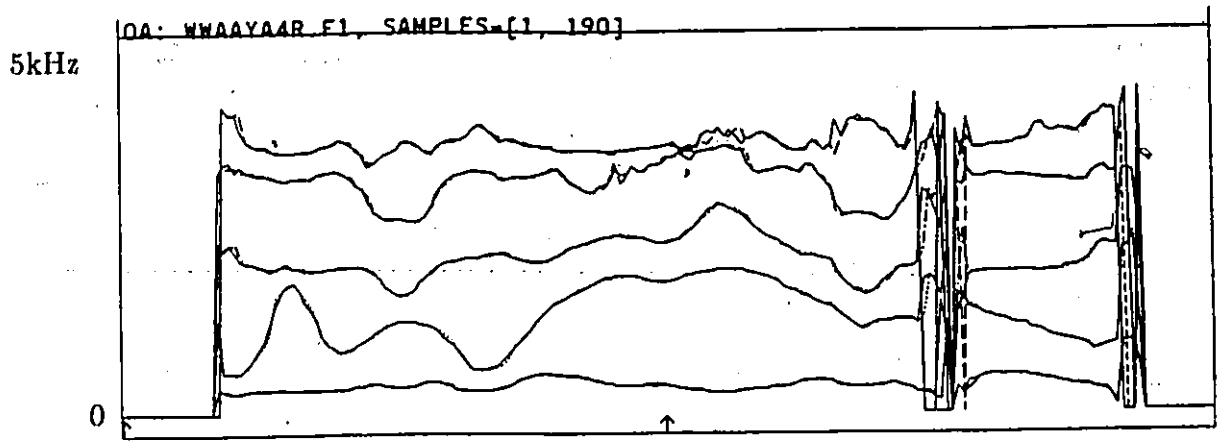

Figure 48: This figure shows the closeness between formant frequencies derived from the pitch-synchronous and pitch-asynchronous analyses (10 ms window with 1 ms overlap). The vertical axis represents the frequency, ranging from 0 Hz at the bottom to 5 kHz at the top. The horizontal axis represents time scale for the duration of the whole sentence. Formants from both analyses are superimposed, with F_1 at the bottom and F_5 at the top. The sentence is "*We were away a year ago.*"

The value of B_1 is prevented from falling below 40 Hz using the formula

$$B_1' = \sqrt{B_1^2 + 1600} \qquad (21)$$

where B_1 is the original bandwidth and B_1' its adjusted value (see Figure 49). Although under-estimation of B_1 is particularly significant during nasals, limiting the value of B_1 was also found to improve the synthesis from pitch-synchronous analyses in a global way (not just for nasals).

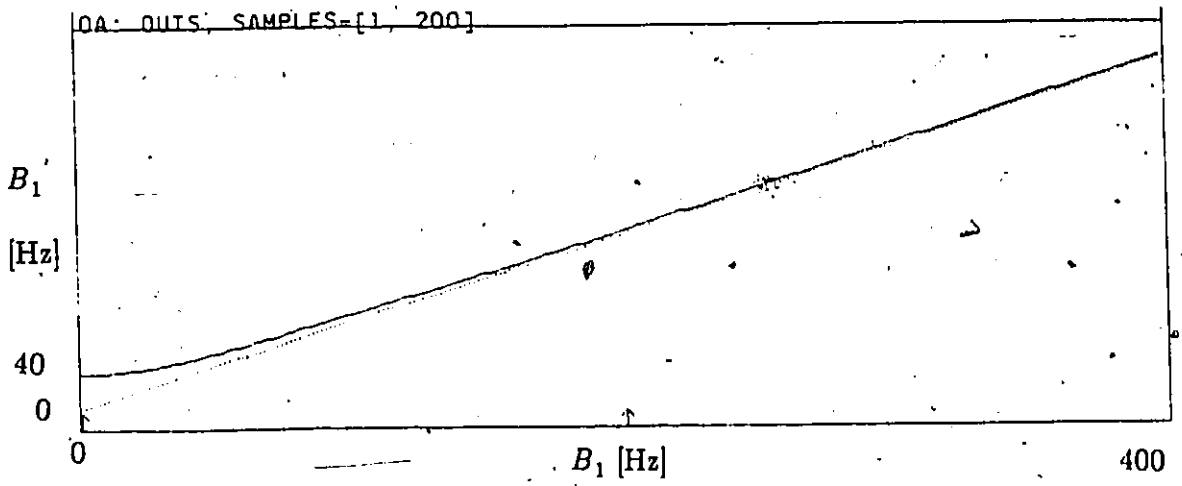

Figure 49: Function used to limit the value of B_1 for synthesis purposes. The dotted line is $B_1' = B_1$.

Figure 50 shows the value of B_1 obtained from both types of analyses.

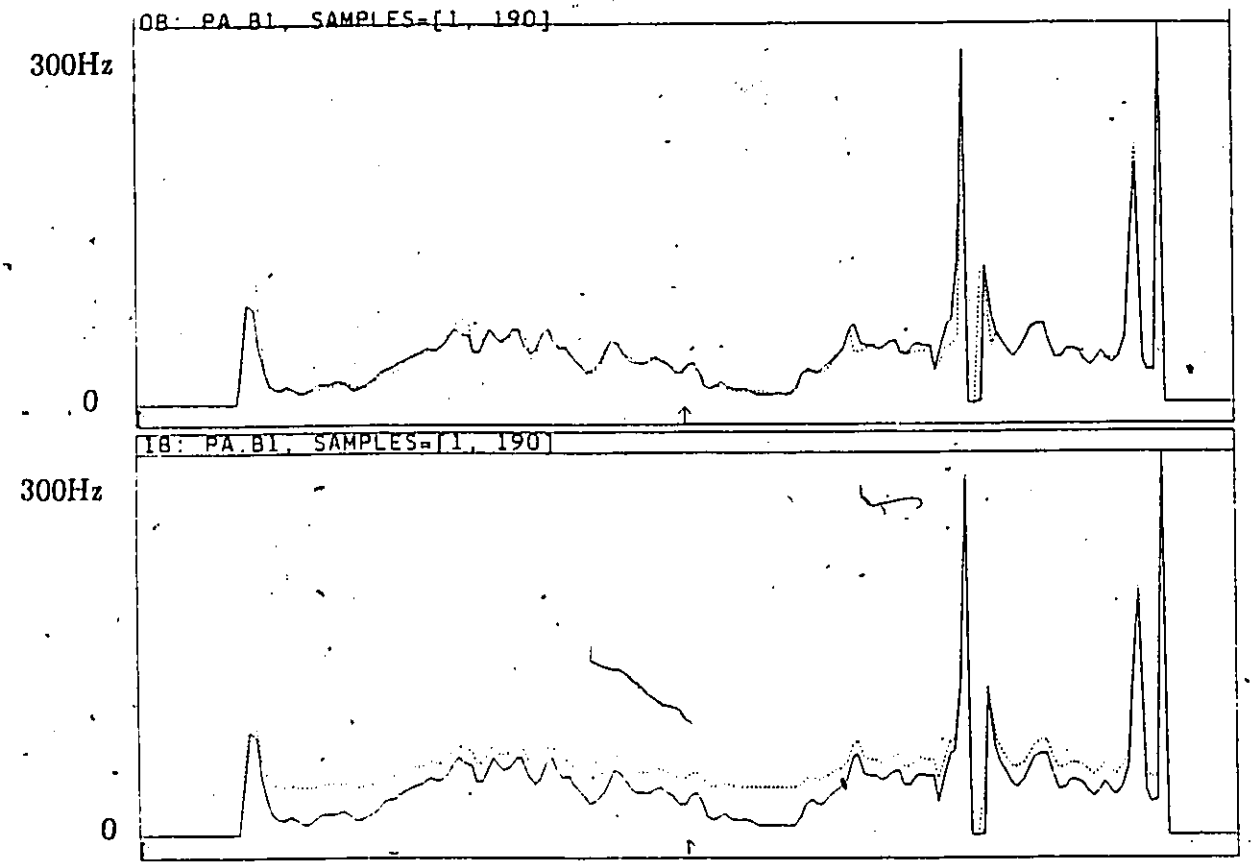

Figure 50: Top window: un-adjusted values of B_1 obtained from the pitch-synchronous analysis (dotted line) and from a pitch-asynchronous analysis (solid line). Bottom window: B_1 from the pitch-asynchronous analysis: un-adjusted (solid line) and adjusted (dotted line). All analyses where done with the LPC covariance method using rectangular windows with a 1 ms overlap.

To explain why the asynchronous and synchronous analyses are comparable, it has been conjectured that LPC in a sense does its own pitch-synchronous analysis. Recall that linear prediction minimizes the square error on the predicted waveform, and therefore concentrates its analysis on the strong parts of the signal. The hypothesis is that LPC concentrates its analysis on the start of each

glottal cycle because this is the strongest part of the speech, and because of this, frame placements within glottal cycles are irrelevant. In an attempt to verify this hypothesis, 2nd order LP analyses were performed on chirps of decaying amplitude — sine waves of linearly changing frequency and exponentially decaying amplitude. These chirps were both 100 samples long (10 ms). LP analyses were performed on two sets of chirps: chirps of decreasing frequency, and chirps of increasing frequency. The results of the LP analyses are shown in Table 1.

start frequency (Hz)	stop frequency (Hz)	pole frequency (Hz)	pole bandwidth (Hz)
1000	950	939	78
950	1000	955	80
1000	990	988	80
990	1000	991	80

Table 1: Results from 2nd LPC analysis on chirps of decaying amplitude.

The results show that to minimize the error power, the pole frequency (the zero of the inverse filter) is not necessarily at the frequency of the strong parts of the signal. This does not mean, however, that the strong parts of the signal do not have a predominant effect on the LP analysis. Comparing the effects on LP analysis of changes in strong parts of the signal to changes in weak parts of the signal suggests that they do. Second-order LP analyses were performed on chirps that were non-uniformly resampled. In one case, the sampling interval for the first 20% of the chirp was unchanged, and the interval was linearly increased to 10% by the time the end of the chirp was reached. This had the effect of linearly increasing the frequency components along the chirp by the same

percentage, and is referred to as "compression at end." In the other case, the sampling pattern was reversed, so that the 10% frequency increase occurred at the start of the chirp. This is called "compression at beginning." Figure 51 shows the chirp in its undistorted form, and Table 2 shows the percentage change in the pole frequency with respect to the undistorted (normal) chirp. These results clearly indicate that the LPC analysis does indeed concentrate on strong parts of the signal.

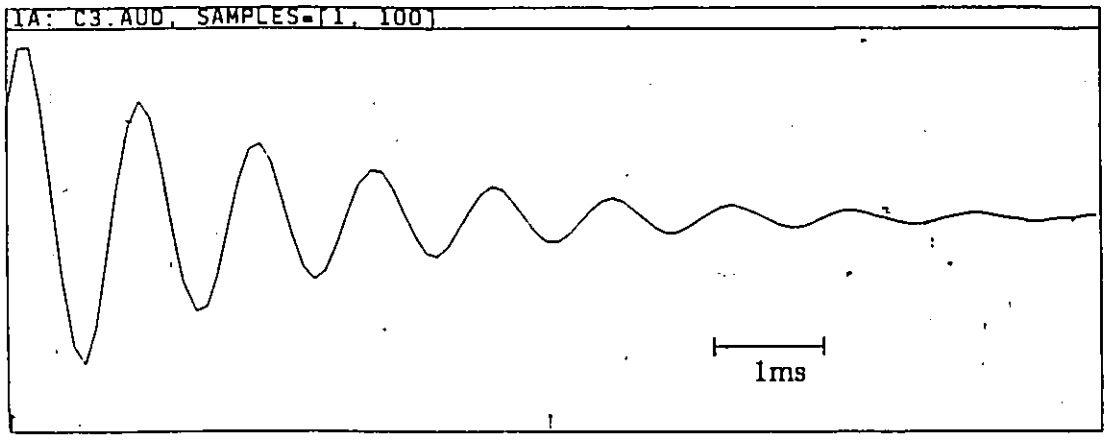

Figure 51: Undistorted chirp used for distortion test: The envelope is $e^{-.04x}$, and the start and stop frequencies are 1000 Hz and 950 Hz respectively.

	pole frequency	percent change
normal	939	—
compression at end	942	.320
compression at beginning	1014	7.99

Table 2: Pole frequencies from second-order LPC analyses performed on normal and non-uniformly resampled chirps.

This resampling test was also performed on four speech frames. The

sampling interval for the first 20% of a frame corresponded to that of the original sampling rate, and then the interval was linearly increased by 10% by the time the end of the frame was reached. This has the effect of linearly increasing the frequency components along a frame by the same percentage. Then, the resampling pattern was reversed, i.e. the 10% frequency increase occured at the beginning of a frame. Tenth order LP analyses were performed on these distorted frames, and the results are plotted in Figure 52, showing the percentage change of formant frequencies relative to the pitch-synchronous analyses. Each line represents a different frame. These results also show a tendency of LP analyses to concentrate on strong parts of the signal. This is not the same thing as saying the frequency returned from an LP analysis corresponds to the frequency of the strong part of a signal (see Table 1).

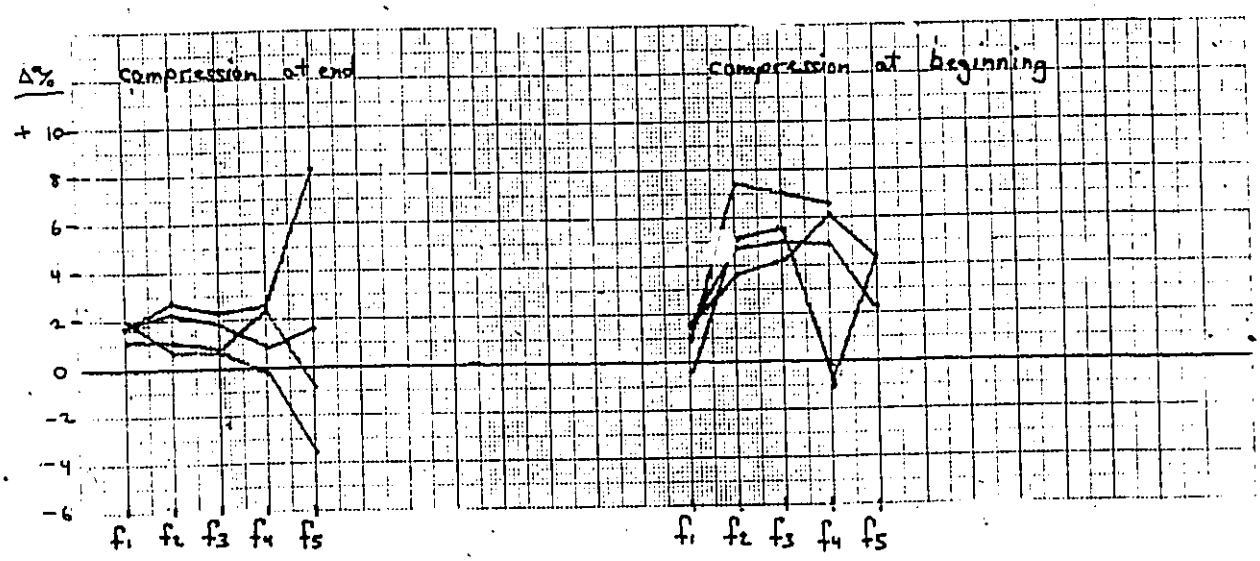

Figure 52: Percentage change in formant frequencies

Another experiment consisted of doing an LPC analysis on different parts of a glottal cycle, and comparing the results. A full glottal cycle analysis

was compared against closed and open glottis analyses. The analyzed speech was a neutral vowal (*schwa*). The open and closed glottis intervals were localized with the glottal air flow signal, shown in the bottom window of Figure 53. Flat segments of this signal indicate zero air flow and hence closed vocal cords. The glottal air flow was derived by taking the residual from the 10th-order LPC analysis and integrating it twice. The small bump in these intervals may have been caused by an incompletely cancelled first formant, or by secondary vocal cord activity.

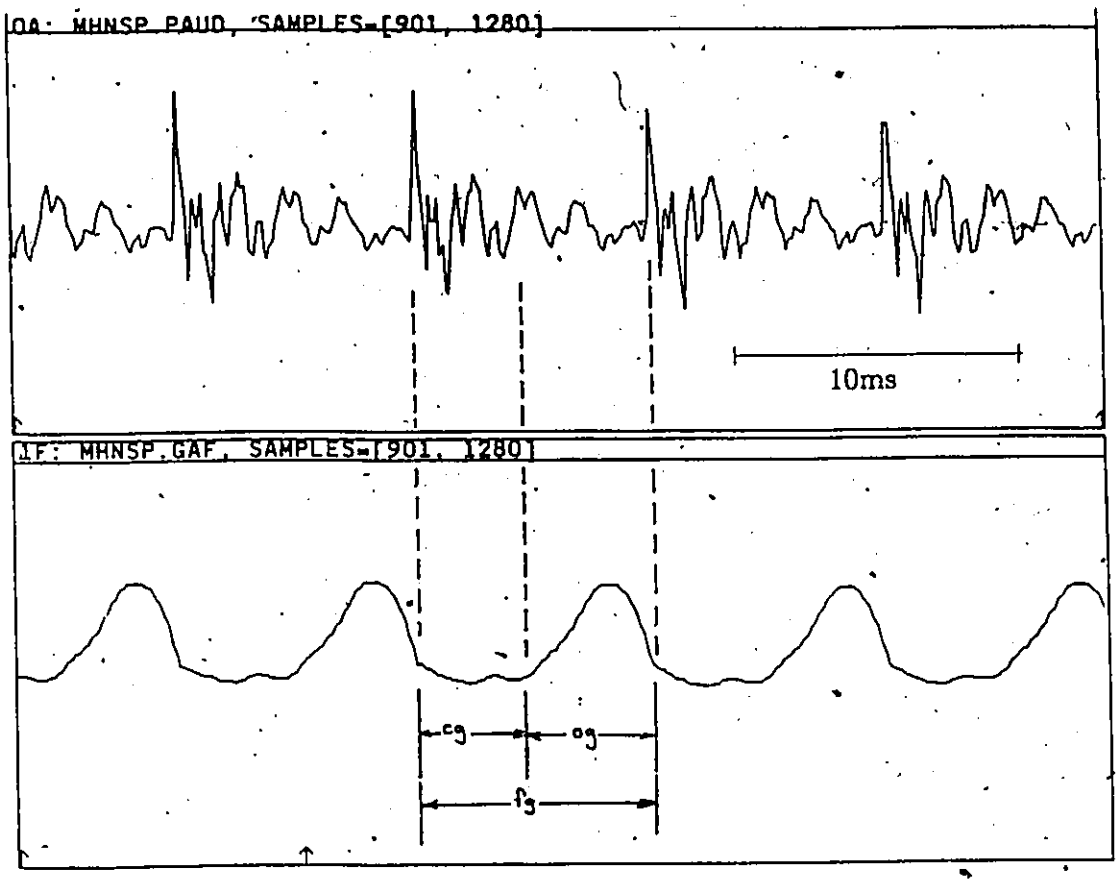

Figure 53: Preemphasized *schwa* vowel (top window) and corresponding glottal air flow (bottom window). "cg" and "og" indicate open glottis and closed glottis intervals, whereas "fg" stands for "full glottal" interval.

Table 3 shows the values of formant frequencies and bandwidths obtained from a 10th order LP analysis of different interval of a glottal cycle. For the particular frame analyzed here, the closed glottis values ("cg") are closer to those of the full glottal cycle ("fg"), as expected. The pole frequencies from the open glottis analysis moved in the opposite direction from that expected from the effect of sub-glottal coupling.

	f1 (Hz)	b1 (Hz)	f2 (Hz)	b2 (Hz)	f3 (Hz)	b3 (Hz)	f4 (Hz)	b4 (Hz)	f5 (Hz)	b5 (Hz)
fg	484	34	1395	56	2191	78	3850	129	—	—
cg	523	25	1410	47	2192	41	3831	123	4656	181
op	550	209	1580	289	2199	222	3978	182	—	—

Table 3: Values of formant frequencies and bandwidths obtained from a 10th order LP analysis of different interval of a glottal cycle. Except for the closed glottis case, no fifth formant was found.

Although the results from this experiment are less convincing, LP analyses seem more sensitive to strong parts of signals than to the weaker parts. It is difficult to assess, however, if the magnitude of this tendency is sufficient to explain the unexpected performance of pitch-asynchronous LPC analyses.

CHAPTER 8

CONCLUSIONS

In the previous chapter we attempted to explain why the performance of a pitch-asynchronous analysis with a fixed 10 ms analysis interval is comparable to that of a pitch-synchronous analysis. It was conjectured that the LPC technique naturally concentrates its analysis on strong parts of the signal, and since these occur in synchrony with the closing of the vocal cords, pitch-asynchronous analysis are effectively pitch-synchronous. Several tests were performed to verify this hypothesis, and the results of these tests support it to various degrees.

The changes to the LPC system significantly improved the naturalness of the synthetic speech. The new speech segmentation program performs well; it has not been observed to make frame classification errors. The U-shaped threshold used to detect peaks in the laryngograph signals increases the reliability of the pitch extraction. Further improvements could consist of an automatic polarity detection and automatic scaling of the laryngograph signal. Both of these steps are done by hand at present, and the addition of this improvement would automate completely the entire analysis and synthesis system.

The temporal structure of the voiced excitation signal is important, and for voiced fricatives, the excitation function needs a degree of aperiodicity

between adjacent glottal cycles to eliminate the perception of buzziness.

Frames containing voiced fricatives are detected automatically using measures of power and B_1. The proper synthesis of these frames contributes significantly to the quality of the synthesis. Even though only short parts of the synthesis are affected, they cannot be neglected, as brief unnatural sounds in otherwise good synthesis are distracting.

Provided a lower bound is imposed on the value of B_1, a re-synthesis from a pitch-asynchronous analysis is found to be comparable in quality to a re-synthesis from a pitch-synchronous analysis.

APPENDIX A: Other modifications

This section explains minor modifications made by the author to the speech segmentation program. The modifications are called "voice onset reflect" and "voiceless onset." Both of these modifications are an attempt to improve the speech synthesis by increasing the accuracy of the excitation gain (or power) parameter.

The *voice onset reflect* algorithm, whose effect is shown in Figure 54, adjusts the length of a last frame before a voiced region so that it has the same pitch-period as the first pitch period detected by the laryngograph. This is reasonable, because the first glottal cycle is usually weak and undetected by the laryngograph. Although this addition to the segmentation process makes a questionable difference on the final synthesis, it is probably, on the average, better than to let the last unvoiced frame before a voiced region be of random length, because the power parameter is still estimated more accurately.

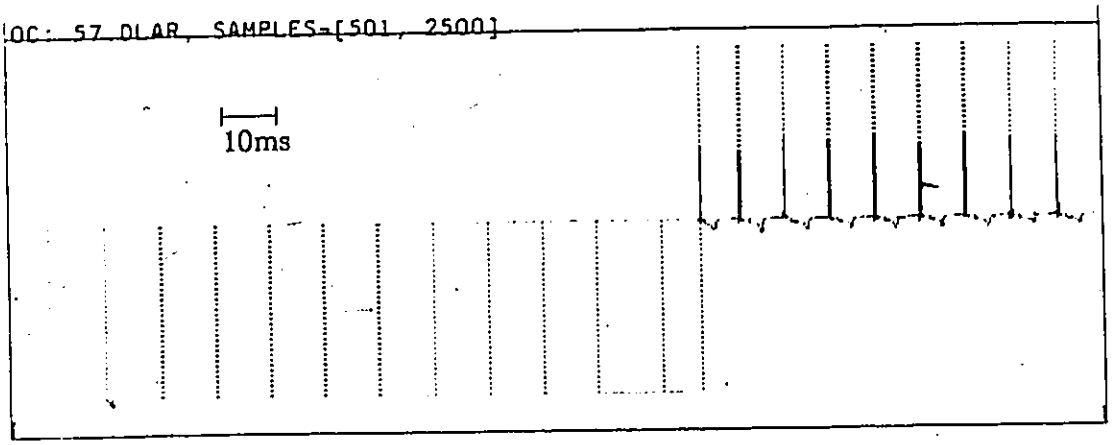

Figure 54: Voice onset reflect. The solid trace is the time-differenced laryngograph signal, and the dotted trace indicates frame boundaries: lines going to the top of the window indicate voiced frames, and lines going down, voiceless frames. The last voiceless frame before a voiced region has the same pitch-period as the first pitch-period detected by the laryngograph.

Voiceless onset attempts to detect more precisely the instant at which voiceless speech begins. In the absence of activity from the vocal cords, the frame sizes are set at 10 ms. This means that the onset of voiceless speech cannot be detected with a resolution of less than 10 ms. As a result of this, errors in the power (gain parameter) for short bursts of voiceless speech may produce degradations in the synthesis.

The author modified the *seg* program to enable it to monitor the short term power in the audio signal in intervals of 1 ms. Therefore, it now detects the onset of voiceless speech with an accuracy of 1 ms, and thus errors in power estimates are greatly reduced (Figure 55).

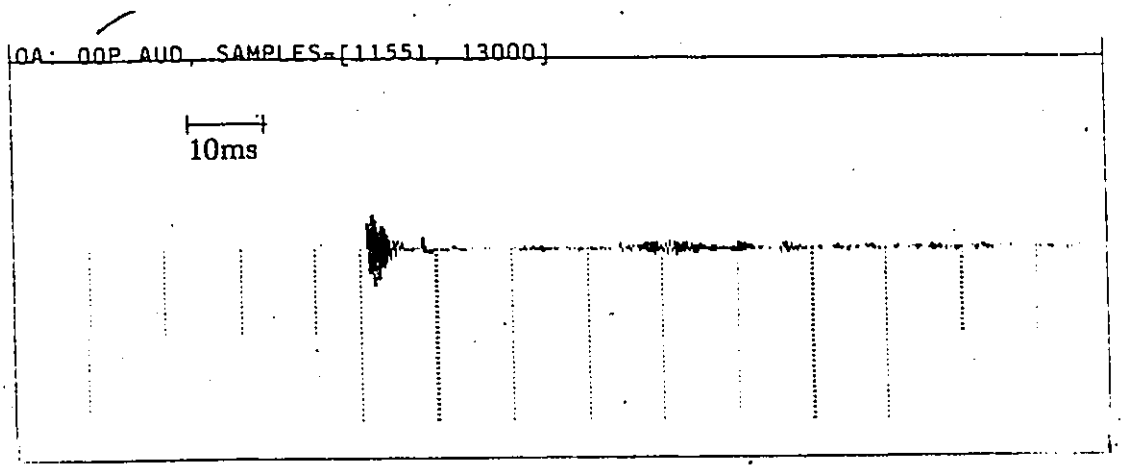

Figure 55: Voiceless onset. The solid trace is the preemphasized acoustic signal for the sound in "Jack," and the dotted vertical bars indicate frame boundaries: lines going half way down the window indicate silent frames, and the line going to the bottom of the window, voiceless frames. Bursts of noise are detected with a 1 ms accuracy.

This modification improves the analysis and synthesis of short bursts of voiceless sounds. To illustrate this, consider the sound, such as the one in "Jack," and assume that it lasts less than 10 ms. If the sound falls right on a 10 ms frame boundary, the power will be split over two frames, and the sound will not be reproduced accurately.

REFERENCES

1. Dudley, H., "The Vocoder," Bell Labs. Record 17, 122-126, 1939.
2. Pisoni, D.B., "Perceptual Evaluation of Voice Response Systems," *Proc. Workshop on Standardization for Speech I/O Technology*, Gaithersburg, MD, March 1982, pp. 185-192.
3. Hunt M.J. & Harvenberg, C.E. "Generation of Controlled Speech Stimuli by Pitch-Synchronous LPC Analysis of Natural Utterances," *Proc. Int. Congress on Acoustics*, Toronto, July 1986, Vol. 1, paper A4-2.
4. Quackenbush, Barnwell III, Clements, *Objective Measures of Speech Quality*, Prentice Hall Signal Processing Series, Englewood Cliffs, New Jersey, 1988.
5. Hunt, M.J.,"Studies of Glottal Excitation using Inverse Filtering and an Electroglottograph," *Proc. XI'th Intl. Congress of Phonetic Sciences*, Tallinn, Estonia, August 1-7, 1987, Vol. 3, pp.23-26.
6. O'Connor, J.D., *Phonetics*, Penguin Books, Harmondsworth, Middlesex, England, 1973.
7. Hunt, M.J., "The Speech Signal," Proc. NATO AGARD Lecture Series No 129, Speech Processing, pp. 2.1-2.12, 1983.
8. Hess, H., *Pitch Determination of Speech Signals: Algorithms and Devices*, Springer-Verlag, Berlin, 1983.
9. Rabiner, L., Cheng, M., Rosenberg, A., & McGonegal, C. "A comparative performance study of several pitch detection algorithms," *IEEE Trans. ASSP-24*, pp. 201-212, 1976.
10. Krishnamurthy, A.K., & Childers, D.G., "Two-channel Speech Analysis," *IEEE Trans. on Acoustics, Speech and Signal Processing*, 1986, Vol 34, pp. 730-743.
11. Holmes, J. N.,"The JSRU Channel Vocoder," *IEE Proc. Communications*, Feb. 1980, Vol 127, part F, pp. 53-60.
12. Pickles, J.O., *Introduction to the Physiology of Hearing*, Academic Press, Inc. (London) Ltd, 1982.
13. Atal, B., & Hanauer, S., "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am. 50*, pp. 180-183, 1971.
14. Holmes, J.N. "Formant Synthesizers: Cascade or Parallel ?" *JSRU Research Report No. 1017*, December, 1982.
15. Rye, J.M. & Holmes, J.N. "A Versatile Parallel-Formant Speech Synthesizer," *JSRU Research Report No. 1016*, November, 1982.

16. Cattermole, K.W. *Principles of Pulse-Code Modulation*, Ilife, England, 1969.

17. Bell Telephone Laboratories, *Transmission Systems for Communication*, 4th ed., 1970.

18. Tomozawa, A. & Kaneko, H., "Companded Delta Modulation for Telephone Transmission," *IEEE Trans. Commun. Technol.*, vol. CT-14, pp. 9-157, February 1968.

19. Klatt, D.H. "The Klattalk Text-to-Speech Conversion System," pp. 1589-1592, *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing ICASSP-82*, Paris, France, May 3-5, 1982.

20. From Text to Speech: The MITalk System, Cambridge University Press, Cambridge, England, 1987.

21. Holmes, J.N., Mattingley, I.G. & Shearme, J.N "Speech Synthesis by Rule," *Lang. & Speech*, 1964, vol. 7, pp. 127-143.

22. O'Shaughnessy, D., *Speech Communication, Human and Machine*, Addison-Wesley Publishing Company, p. 377, 1987.

23. Markel, J.D. and Gray, A.H.Jr., *Linear Prediction of Speech*, Springer-Verlag, Berling, 1976.

24. Clasen, R.J., "Numerical Methods for Inverting Positive Definite Matrices," Rand Corporation, Santa Monica, California, AD637-930, 1966.

25. Hunt, M.J., Zwierzynski, D., & Carr, R.C., "Issues in High Quality LPC Analysis and Synthesis," *Proc. European Conference on Speech Communication and Technology*, September 26-28, Paris, France, 1989.

26. Un, C.K. & Magill, D.T., "The Residual-Excited Linear Prediction Vocoder with Transmission Rate below 9.6 Kbits/s," *IEEE Trans. on Comm.*, Dec. 1985, Vol. COM-23, pp. 1466-1474.

27. Atal, B., & Remde, J., "A new model of LPC excitation for producing natural-sounding speech at low bit rates," *Proc. IEEE Int. Conf. ASSP*, pp. 614-617, 1982.

28. Rosenberg, A., "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am. 73*, S1, S5, 1983.

29. Minsky, M. and Papert, S., "Perceptrons: An Introduction to Computational Geometry," MIT Press, 1969.

30. Lippmann, Richard P., "An introduction to Computing with Neural Nets," *IEEE ASSP magazine*, April 1987, Vol. 4, Number 2

31. Deterding, D. H., "Pitch Synchronous Linear Prediction," Cambridge Papers in Phonetics and Experimental Linguistics, Vol 5, 1986, Department of Linguistics, University of Cambridge