Fall 1993

Acoustic Correlates of Vowel Perception as Determined from Synthesis Experiments With Multi-Tone Stimuli

Zhongjiang Zhang
Old Dominion University

ACOUSTIC CORRELATES OF VOWEL PERCEPTION AS DETERMINED FROM SYNTHESIS EXPERIMENTS WITH MULTI-TONE STIMULI

Zhongjiang Zhang

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

MASTER OF SCIENCE

ELECTRICAL ENGINEERING

OLD DOMINION UNIVERSITY
December 1993

Approved By:

Stephen A. Zahorian (Director)

Peter L. Silsbee

S. Nandkumar

Zaki B. Nossair

ACOUSTIC CORRELATES OF VOWEL PERCEPTION AS DETERMINED FROM SYNTHESIS EXPERIMENTS WITH MULTI-TONE STIMULI

Zhongjiang Zhang

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

MASTER OF SCIENCE

ELECTRICAL ENGINEERING

OLD DOMINION UNIVERSITY
December 1993

ABSTRACT

ACOUSTIC CORRELATES OF VOWEL PERCEPTION AS DETERMINED FROM SYNTHESIS EXPERIMENTS WITH MULTI-TONE STIMULI

Zhongjiang Zhang
Old Dominion University
December 1993
Director: Dr. Stephen A. Zahorian

An essential requirement of speech signal processing is to extract information (features or parameters) from the speech signal which encode the information carried by the signal. The objective of this thesis work was to examine and evaluate two feature sets as acoustic correlates for vowel perception. They are formants and DCTCs. Formants are the frequencies of spectral peaks of the speech signal. DCTCs are the Discrete Cosine Transform Coefficients of the magnitude spectrum and are thus features which encode the global spectral shape of speech signal.

There are different opinions regarding which feature set is a more accurate representation for vowels. In fact the parameters most useful for automatic speech classification may not be good acoustic correlates for the perception of speech. Based on the results of Zahorian and Jagharghi (1990, 1993), we initially hypothesized that global spectral shape cues are more important to phonological perception of vowels than are formant frequency cues.

The higher-level objective of the study was to determine a feature set based on certain aspects of both formant and global spectral shape theory, which would be good acoustic correlates of vowel perception. We developed and investigated a new algorithm to compute the DCTCs which represents the spectral shape of the envelope of the speech spectrum. It requires only about 10 percent of the Fourier Transform magnitude components as compared to the DCTCs computed by Zahorian and Jagharghi.

Experiments conducted in this thesis work support the hypothesis that formants are insufficient acoustic correlates for vowel perception and that some type of global spectral features are required. The original DCTC features were also found to be lacking as acoustic correlates of perception. However, a modified DCTC computation was formulated which results in more perceptually significant features. These new features also improve automatic vowel classification of noisy speech. Topics for further study are suggested.

ACKNOWLEDGEMENT

I would like to extend great thanks to my advisor, Dr. Stephen A. Zahorian, for his support and guidance. He led me into this field from a novice background. This thesis would not have been possible without his knowledge, wisdom, and direction.

I would like to thank Dr. Zaki B. Nossair for his helpful advise throughout the research work. I would also like to thank the additional members of the committee, Dr. Peter L. Silsbee and Dr. S. Nandkumar, for their time and advice. I also thank all coworkers in the lab for their assistance.

Special thanks to my parents and my family numbers for their understanding, patience and help.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENT i

TABLE OF CONTENTS ii

LIST OF TABLES v

LIST OF FIGURES vi

INTRODUCTION 1

1.1 Outline of Speech Processing in This Study 4

1.2 Sinusoidal Speech Synthesis Model 5

1.3 Organization of Thesis 8

SPEECH SIGNALS AND FEATURES:
AN EXPERIMENT TO EVALUATE THE FEATURES 9

2.1 The Speech Signal 9

2.2 Features of the Speech Signal 15

2.2.1 Formants 16

2.2.2 DCTCs 20

2.2.3 Pitch 23

2.3 Data Recording 23

2.4 Format of Experiments 24

EXPERIMENTS ON FORMANTS VERSUS DCTCs 27

3.1 Synthesis with Sinusoids Based on Formant and DCTCs 29

3.1.1 Uniformly Spaced Sinusoids 30

3.1.2 Bark Space Sinusoids to Preserve Spectral Envelope 32

3.1.3 Bark Space Sinusoids to Preserve Global Spectral Shape 34

3.1.4 Sinusoids to Preserve Formant Frequencies 35

3.2 Listening Experiment 37

3.3 Results 37

3.4 Conclusion from this Experiment 40

REFINING ACOUSTIC CORRELATES FOR VOWEL PERCEPTION 42

4.1 DCTC Peak Algorithm 42

4.2 Synthesis Experiment 46

4.2.1 Original Vowel and Repetition 47

4.2.2 Varying Numbers of Harmonic Peaks 49

4.2.3 Formant Harmonics 50

4.2.4 Bark Spaced Harmonics 52

4.2.5 The Largest Peaks 53

4.3 Listening Experiment (II) 54

4.4 Results 54

4.5 AXB Experiment 58

CLASSIFICATION 61

5.1 Classifier 62

5.2 The Computation of Normal DCTC and
DCTC Envelope Coefficients 65

5.2.1 The Common Variables 65

5.2.2 The Method of Selecting Peaks 67

5.2.3 Methods for Computing DCTCs Which Encode
the Envelope Spectrum 70

5.3 Primary Experiment 75

5.4 Experiment Based on Noisy Speech 81

CONCLUSION 86

BIBLIOGRAPHY 90

APPENDIX A.
Confusion Matrix of Listening Experiments for Chapter 4 92

APPENDIX B.
Classification Results for Chapter 5 98

LIST OF TABLES

Table		Page
3-1	Formants of five vowels for male and female speakers	28
3-2	Confusion matrix for original tokens	39
3-3	Confusion matrix for case 1, token synthesized with 30+ uniformly spaced sinusiods to match both spectral envelope and global spectral shape	39
3-4	Confusion matrix for case 2, token synthesized with 16 nonuniformly spaced sinusiods to match spectral envelope	39
3-5	Confusion matrix for case 3, token synthesized with 16 nonuniformly spaced sinusiods to match global spectral shape	40
3-6	Confusion matrix for case 4, token synthesized with 3 nonuniformly spaced sinusiods to match formant frequencies and amplitudes	40
4-1	Summary of synthesis conditions	48
4-2	All possible formants values for 10 American English and three speaker type groups	51
4-3	Results of the experiment	57
4-4	List of conditions for AXB experiment	59
4-5	List of results of AXB experiment	59
5-1	Number of vowel tokens in training set and testing sets	76

LIST OF FIGURES

Figure		Page
2-1	Spectrogram of a speech signal "Beep"	11
2-2	Speech signal "Beep" in time domain	12
2-3	Vowel /iy/ extracted from speech signal "Beep" and expanded in time domain	13
2-4	FFT spectrum of vowel /iy/	15
2-5(a)	Fomants of a male speaker for vowel /iy/	17
2-5(a)	Fomants of a male speaker for vowel /ih/	18
2-5(a)	Fomants of a female speaker for vowel /iy/	18
2-6	The listening experiment environment screen	25
2-7	AXB experiment response screen	26
3-1	Illustration of frequencies and amplitudes of uniformly spaced sinusoids	31
3-2	DCTC spectrum for original and synthesized speech using uniformly spaced sinusoids	31
3-3	Illustration of frequencies and amplitudes of Bark spaced sinusoids to preserve spectral envelope	33
3-4	DCTC spectrum for original and synthesized speech of Bark spaced sinusoids which preserve the spectral envelope	33

3-5 Frequencies and amplitudes of Bark spaced sinusoids which preserve global spectral shape 34

3-6 DCTC spectrum for original and synthesized speech of Bark spaced sinusoids which preserve global spectral shape 35

3-7 Frequencies and amplitudes of sinusoids which preserve formant frequencies 36

3-8 DCTC spectrum of original and synthesized speech from sinusoids which preserve formant frequencies 36

3-9 Experimental results 38

4-1 Comparison of normal DCTCs spectrum and peak DCTCs spectrum 45

4-2 Illustration of 8 largest DCTCs spectral peaks for /aa/.......... 50

4-3 Formant frequencies plus side peaks 52

4-4 Bark scale peaks selected as sinusoids 53

4-5 Four largest peaks plus side peaks as sinusoids 55

4-6 Eight peak sinusoidal speech, without the four largest peaks 55

4-7 Bar graph of experimental results 56

5-1 Illustration of the effect of the degree of warping on vowel classification for several envelope DCTC computation methods 66

5-2 Illustration of spectrum from normal DCTC and original FFT spectrum 68

5-3 Illustration of the envelope DCTCs spectrum with computations based on harmonically related peaks (method 1 in text) 69

5-4 Illustration of DCTC spectrum computed using the harmonic peaks + linear interpolation method 72

5-5 Illustration of spectrum computed from 15 envelope DCTCs computed using largest peaks in Bark_spaced + linear interpolation method (method 4 in text) 73

5-6 Illustration of spectrum computed from 16 envelope DCTCs computed using the uniform Bark-spaced peaks + linear interpolation method (method 6 in text) 74

5-7 Automatic vowel classification results (13 vowels) for six envelope DCTC computation methods, as a function of the number of DCTCs used 75

5-8 Automatic vowel classification results obtained with normal DCTCs and three types of envelope DCTCs 77

5-9 Illustration of the effect of varying the number of Bark-spaced peaks used for the largest peaks in Bark-spaced peaks + linear interpolation method (method 4 in text) 78

5-10 Illustration of the effect of varying the number of Bark-spaced peaks used for the uniform Bark-spaced peaks + linear interpolation method (method 6 in text) 79

5-11 Illustration of clean speech and noisy speech (SNR = 5 dB) 81

5-12 Vowel classification results for normal DCTCs and envelope DCTCs, using one frame of speech, at various signal-to-noise ratios 83

5-13 Vowel classification results obtained with DCTC trajectories for both normal DCTCs and envelope DCTCs for clean speech and noisy speech (SNR = 0 dB) with multi-frame 84

5-14 Illustration of normal DCTC spectrum and envelope DCTC spectrum for a natural speaker 85

CHAPTER ONE

Introduction

Speech is the most important communication modality for humans. Communication between machines and humans using speech would also be very advantageous. One of the major objectives of speech signal processing is to study this method of communication between humans and machines. Two main branches of study in this field are automatic speech recognition (ASR) and speech production or synthesis. In both cases an essential requirement is to extract information from the speech signal. That is, parameters or features must be computed which represent the information in the speech. These parameters are thus acoustic correlates of perceptual "units," called phonemes, such as vowels. One of the fundamental problems in speech processing is that there is a great deal of variability in the acoustic signal for the same phoneme, due to speaker, phonetic context, etc. Therefore it is very difficult to determine a set of acoustic correlates which are closely linked to phonetic classes.

The objective of this thesis work was to examine and evaluate two feature sets as acoustic correlates for vowel perception. The first set was formants, that is the frequency of spectral peaks of the speech signal. Formants have traditionally been favored by speech scientists because a large amount of speech information is contained in the first three formants, and formants are correlated with vowel perception. Features which encode the global spectral shape are another representation which can be used for vowels. The global spectral shape features which were examined in this work were based on the Discrete Cosine Transform Coefficients of the magnitude spectrum and are thus referred to as DCTCs.

Several different studies have presented different opinions regarding which feature set, formants or DCTCs, is a more accurate representation of vowels. A recent study (Zahorian and Jagharghi,1993) investigated in detail and compared the two sets of spectral features for automatic classification of vowels. It was shown that performance based on global spectral shape is superior to that based on formants. They therefore concluded that spectral shape features are a more complete set of acoustic correlates for vowel identification than are formants.

However, the parameters most useful for automatic speech classification may not be good acoustic correlates for the perception of speech. That is, a set of speech features may work well for automatic classification but may not predict the perception of speech. More specifically, it may be possible to synthesize two tokens of speech with the same values of these parameters, but which are perceived differently, or it may be possible to synthesize two tokens of speech with different values of these parameters, but which are perceived the same. Based on the results of Zahorian and Jagharghi, we initially hypothesized that global spectral shape cues are more important to phonological perception of vowels than are formant frequency cues. Several experiments were conducted to investigate this hypothesis.

The higher-level objective of the study was to determine a feature set based on certain aspects of both formant and global spectral theory, which would be good acoustic correlates of vowel perception. We implicitly assume the theory of acoustic-phonetics, which claims that acoustic correlates exist for phonemes(1). During the course of pursuing this higher-level objective, we developed and investigated a new algorithm to compute the DCTCs which represent the spectral shape of the envelope of the speech spectrum. It requires only about 10 percent of the Fourier Transform magnitude components as compared to the DCTCs computed by Zahorian and Jagharghi.

Two general types of experiments were conducted in this investigation. In the first case, speech was synthesized as a sum of sinusoids, so as to either preserve or modify an assumed set of acoustic correlates, and perceptual listening tests were performed to examine the degree to which those correlates corresponded to vowel perception. A series of these experiments were used to refine the methods used for defining the correlates. In addition, automatic vowel classification experiments

(1). Some researchers have argued that such correlates do not exist, and that features are highly context dependent.

were also used to determine the extent to which the refined features improved automatic vowel classification accuracy.

1.1 Outline of Speech Processing in This Study

Speech, the acoustic signal generated by a human speakers, is a nonstationary process; the instantaneous position of the vocal tract changes with time. Therefore, the speech signal is difficult to describe in a stationary form. The features computed encode the information in the speech signal. These features can also be used by a machine to recognize the speech, or as model parameters for speech synthesis, provided the features can be automatically computed by a machine algorithm.

In order to perform our experimental work, we first recorded speech vowel sounds. The details of the procedure are discussed in Chapter 2, section 3. These vowel sounds were produced in isolation and each speaker was asked to "hold" the vowel for at least one second. Speech features were extracted in non real-time from binary files of these sounds. Speech was then synthesized with a sum-of-sinusoids synthesizer to generate synthesized speech for several conditions. These conditions were based on different feature sets and were used to evaluate the degree to which the features were correlated with perception of the sounds. The detailed conditions are discussed in Chapters 3, 4, and 5. Two kinds of listening experiments were used to examine the human perception of the synthesized speech.

In the first case, commonly called the forced choice paradigm, the listener hears the speech sound and then attempts to identify it from a closed set of possibilities. This experiment is discussed in more detail in Chapters 3 and 4. The second type of listening experiment is called the AXB paradigm. In this comparison test, the listener hears three speech sounds in rapid succession. The middle sound, X, is the target or control sound. The listener must respond as to whether the first (sound A) or third (sound B) is more similar to the X sound. For our experiments, all three sounds in a group were of the same vowel. Usually the X sound was the original vowel, and the A and B sounds were the vowel synthesized with two competing synthesis conditions. More details of this type of listening experiment type are given in Chapter 4. The description, results, and interpretation of the vowel classification experiments is the topic of Chapter 5.

1.2 Sinusoidal Speech Synthesis Model

There are many different models for speech synthesis. Articulatory synthesis is one method which can be used. It attempts to faithfully model the mechanical motions of the articulators and the resulting distribution of volume velocity and sound pressure in the lungs, larynx, and vocal and nasal tracts (Flanagan, Ishizaka, and Shiply, 1975). This method requires extensive computations, and the resultant speech output cannot be specified with sufficient precision for psychophysical experimentation. Another method is formant synthesis. It is based on an acoustic theory of speech production (Fant,1960) and an approximation to the speech waveform by a simple set of rules formulated in the acoustic space. Two general configurations of this modeling are cascade and parallel. Parallel formant synthesizers (Lawrence, 1953; Holmes, 1973) model the transfer function of the vocal tract using several stages connected in parallel. Each formant resonator is preceded by an amplitude control that determines the relative amplitude of a spectral peak (formant) in the output spectrum of the speech. The cascade form connects the formant resonators in a series or cascade fashion (Fant,1959; Klatt, 1972). In contrast to the parallel form, it does not need individual amplitude controls for each formant. A flexible software formant synthesizer, which includes options for both basic forms plus combinations of these, has been developed by Klatt (1980). The software has also been widely distributed.

Rather than make use of either of the above mentioned synthesizers, a sinusoidal model has been used in this study. Not only is the model simpler than these other models, it provides the required flexibility for the tests of this research. The sinusoidal model is based on a Fourier series representation of the speech signal. That is, the frequency components in the speech signal are used to adjust the various aspects of the speech signal.

The definition of the sinusoidal wave function is

$$S(n) = \sum_{i=1}^{N_f} A[i]\, Sin(2\pi \cdot \frac{f[i]}{f_s} \cdot n) \tag{1.1}$$

where A[i] and f[i] are the amplitude and frequency of the sinusoidal synthesizer respectively. The N_f represents the number of components used in the synthesizer and the index i is the i^{th} sinusoid in the synthesizer. A variable L is the number of samples of data in the signal. For this study we used a 16 kHz sampling rate and one second long of synthesis vowel segments (i.e. 16000 samples). (For listening experiments, the middle .56 second section (8960 samples) was used.)

From the equations we can see the advantages of this sinusoidal model. We can use any desired amplitude and frequency (and phase) with this model. That is, we can use acoustic correlates such as formants, which directly imply certain frequencies and amplitudes, or DCTC coefficients (which can be transformed to rebuild a spectrum) to adjust the frequencies and amplitudes for synthesis. Therefore, we can use either formant frequency components or DCTC-derived spectral components for synthesis control. We can also control the number of sinusoids used. This synthesizer can thus be used to either preserve the formant frequencies and amplitudes very precisely, or to preserve the spectral shape very precisely, thus enabling perceptual tests to be made to compare the two types of features.

1.3 Organization of Thesis

The main topic of this thesis is to investigate and formulate acoustic correlates for vowel perception. In Chapter 2 the methodologies used in this study are explained. It contains three sections; a section describing the computation of acoustic parameters (formants, DCTCs, and pitch); a section describing the sinusoidal vocoder; and a section describing the format of the experiments. The initial hypothesis, that acoustically-invariant cues for speech perception are more closely related to global spectral shape than to formant frequencies, is examined in Chapter 3, using experiments based on the sinusoidal synthesizer. Chapter 4 addresses a modified method for computing spectral shape features, and describes experiments used to evaluate these new features. The automatic classification experiments used to verify and refine the new feature set are described in Chapter 5. The conclusion of this study and suggestions for future study are discussed in Chapter 6.

CHAPTER TWO

Speech Signals and Features: An Experiment to Evaluate the Features

This chapter presents a brief description of speech signals, including their physical aspects and the mechanism for human speech production. It describes the general properties of the signal resulting from the production process and describes a methodology for quantitative analysis of the signal. It also presents the format of the listening experiments used in this study to evaluate the perceptual importance of the features used.

2.1 The Speech Signal

One type of speech sound is produced by a continuous flow of air through the vocal tract. It requires the coordination of effort between the lungs, trachea, and the larynx. The force of the air comes from the lungs through the trachea to the larynx and forms the energy for producing the speech sound. There are two flaps of tissue, called vocal cords, situated in the larynx. When air passes between them, they adjust their position and tension to periodically interrupt the air stream and produce a voiced sound. Another type of speech sound is called unvoiced because it is produced by air turbulence not associated with the vocal cords.

Everyone's vocal cords has a characteristic frequency, called the fundamental frequency of voicing, or F0, which depends on the tension of the vocal cords. The speaker can alter the fundamental frequency by controlling the tension of the vocal cords. Note that although the terms fundamental frequency of voicing and pitch are often used interchangeably, actually pitch is a perceptual quality, which is mainly affected by fundamental frequency of voicing. Normally, a man's fundamental frequency is lower than that of a woman or child. The speaker also can alter the shape of the vocal and nasal cavities by moving the tongue and reshaping the lips to produce vowel sounds (such as a, e, i, o, u, and other phonemes). The nasal cavity is also important for some sounds such as the nasal phonemes (m, n, and *n*). There are other sounds such as aspirants (h), fricatives (s,z), and so on. One distinct category of sounds is stops (b, d, g, p, t, k), which are formed by first blocking (i.e., stopping) the air flow in the vocal tract and then releasing it in a burst.

From these descriptions of the phonemes we can see that speech is comprised of some complex sounds. The complexity is evident visually in a speech spectrogram, as shown in figure 2.1.

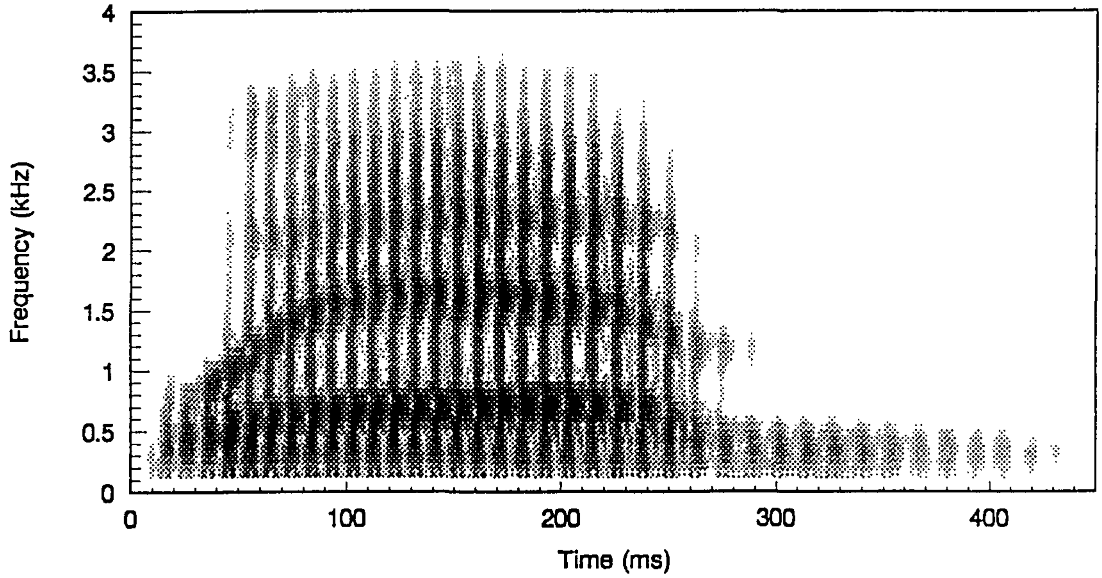

Figure 2.1 Spectrogram of a speech signal "Beep".

The vertical axis on the figure represents frequency, the horizontal axis time, and the density of the trace indicates the energy of the speech at that time and frequency. The dark horizontal bands which are evident in the spectrogram represent the formants, as mentioned in previously. Figure 2.2 depicts the speech signal in the time domain. The vertical axis is amplitude and the horizontal axis is time. This representation is called the acoustic waveform of the speech signal, since it depicts acoustic pressure variations as a function of time.

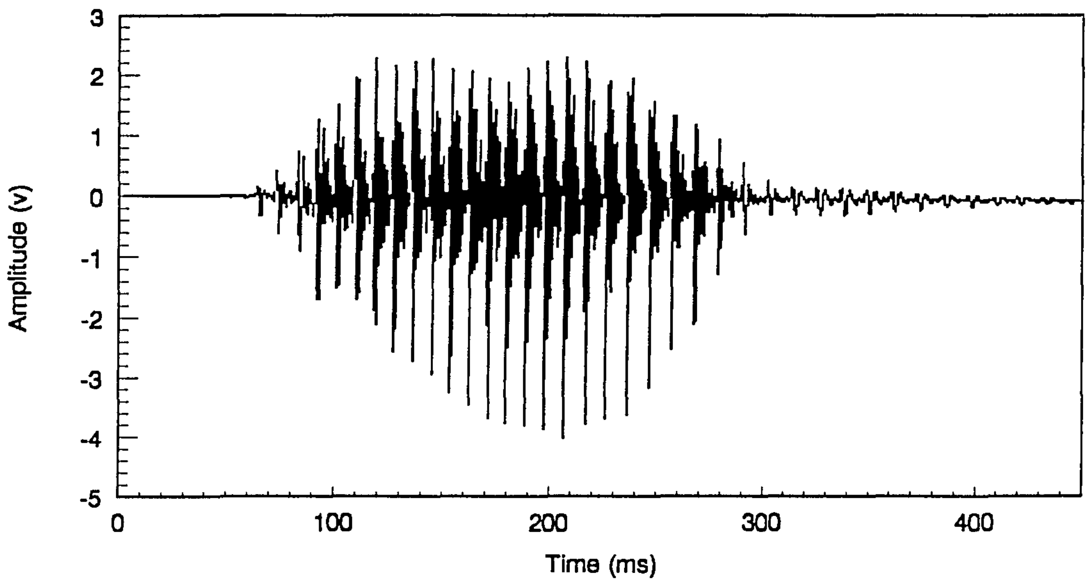

Figure 2.2 Speech signal "Beep" in time domain.

Speech sounds also can be generated by a machine. The resultant sound is called synthesized speech. To illustrate this process, we first expand a portion of the signal in figure 2.2 and depict this small portion in figure 2.3.

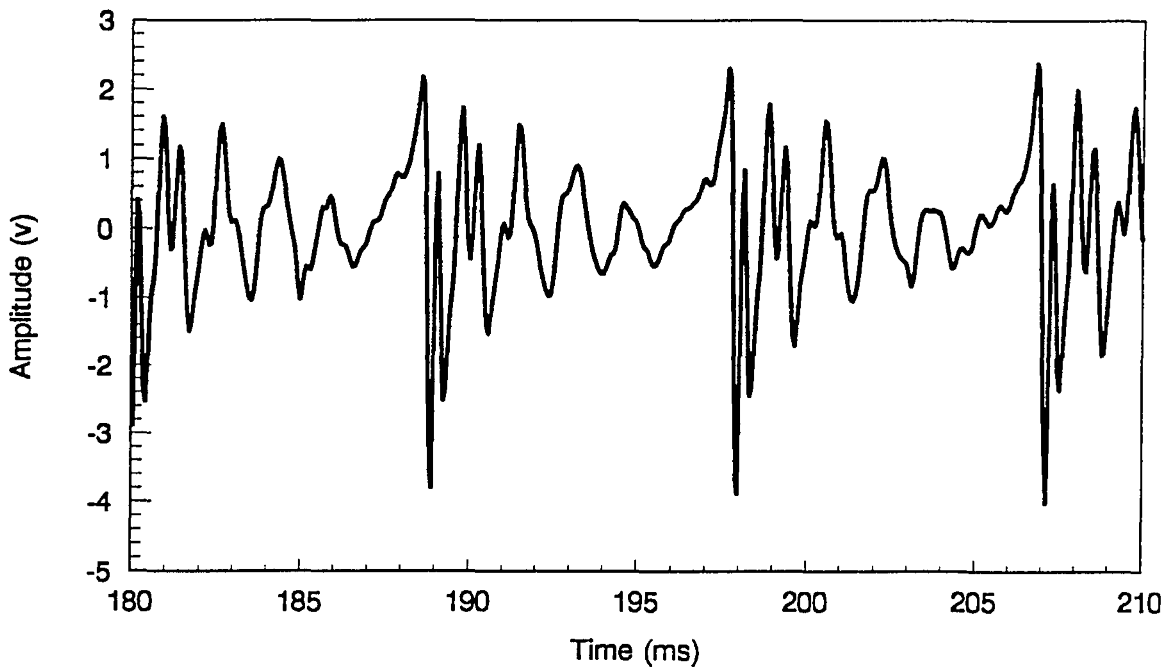

Figure 2.3 Vowel /iy/ extracted from speech signal "Beep" and expanded in time domain.

The details of the signal can be seen well in the time domain. The quasi-periodic signal shown is part of the vowel /iy/. The duration of the signal from one peak to the next peak is about 9 (ms), thus implying a period of 9 ms, or a fundamental frequency of about 110 Hz. From a signal theory point of view the signal can be approximated as a sum of sinusoidal waveforms. This approach leads to a synthesis model called a sinusoidal synthesizer, as discussed in Chapter 1. The parameters of the sinusoid can be obtained from the analysis of the speech signal in the frequency domain -- that is, spectral analysis. Figure 2.4 is the frequency domain representation of the signal. This figure depicts the amplitudes of the frequency components which comprise the speech signal. Note that the phase information, which is known to be relatively unimportant for speech perception, has been discarded.

Although both figure 2.1 and figure 2.4 depict the speech signal, certain characteristics of the signal are much more apparent in one representation whereas other characteristics are more apparent in the other representation. For example, figure 2.1 clearly illustrates the formant characteristics whereas figure 2.4 more clearly illustrates the global spectral shape of the signal. Therefore, different methods of representation of the speech signal are beneficial for different applications. The formants and global spectral shapes representations of the speech signal are discussed in the next section.

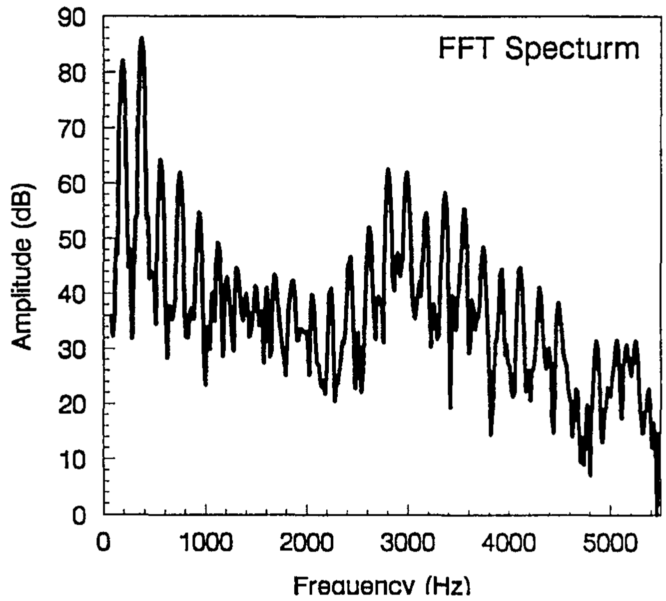

Figure 2.4 FFT spectrum of vowel /iy/.

2.2 Features of the Speech Signal

Both for engineering applications and from a speech science point of view, it is advantageous to represent the speech signal with a compact set of parameters called features. In the previous sections we mentioned that we will examine two such candidate feature sets for representing the speech in acoustic space. They are formants and global spectral shape features. Formants are traditionally considered, at least among speech scientists, to be primary acoustic cues to vowel identity. Global spectral shape features, in particular Discrete Cosine Transform Coefficients, form a more complete representation, but have the disadvantage of requiring more parameters.

2.2.1 Formants

The speech acoustic signal is transmitted through the vocal tract. The vocal source is a wideband excitation. The vocal tract acts as a filter, allowing only certain frequencies to be present in the sounds as they are released from the mouth. If this filter is modeled as a linear system with poles and zeros, the formants correspond to resonances of the vocal tract (i.e., the poles) and result from constrictions in various positions of the vocal tract. Formants are denoted as F1, F2, F3 etc., in order of increasing frequency. Each vowel has a different pattern of resonances than the others. Therefore, formants can be used to characterize vowels by speech scientists. However, not all speakers have the same formant values for the same vowel. Typically, female and child voices have higher frequency formants than do males. Figures 2.5(a)-(c) illustrate these characteristics. Figure 2.5(a) and (b) are for the same speaker with different vowels. Figure 2.5(b) and (c) are for the same vowel with different speakers. In addition to the inter-speaker variations, the shape of vocal tract of a speaker may vary with time, weather, and other factors. Another practical difficulty with using formants as features, at least for applications such as automatic speech recognition, is that it is often extremely difficult to automatically identify them.

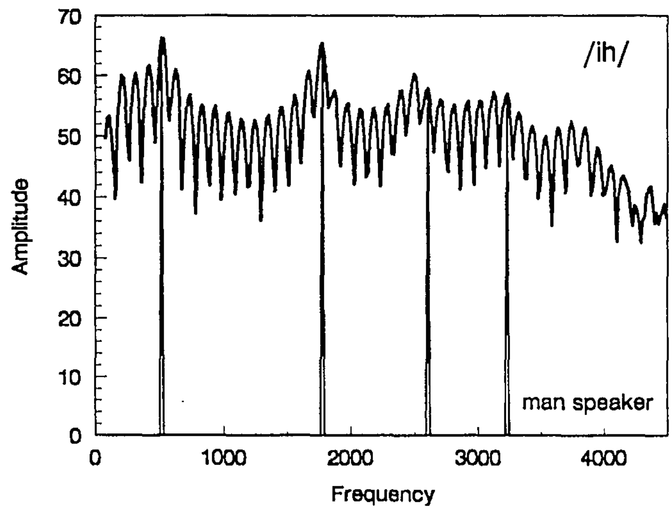

Figure 2.5(a) Formants of a male speaker for vowel /ih/.

Therefore, the performance of an automatic speech recognizer based on formants is obviously degraded. However, on the positive side, formants do carry considerable speech information with only three features, and they are thus considered to be an important feature set.

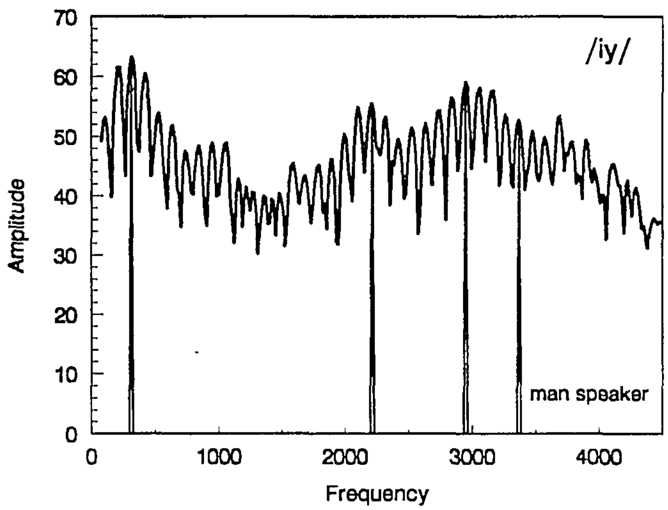

Figure 2.5(b) Formants of a male speaker for vowel /iy/.

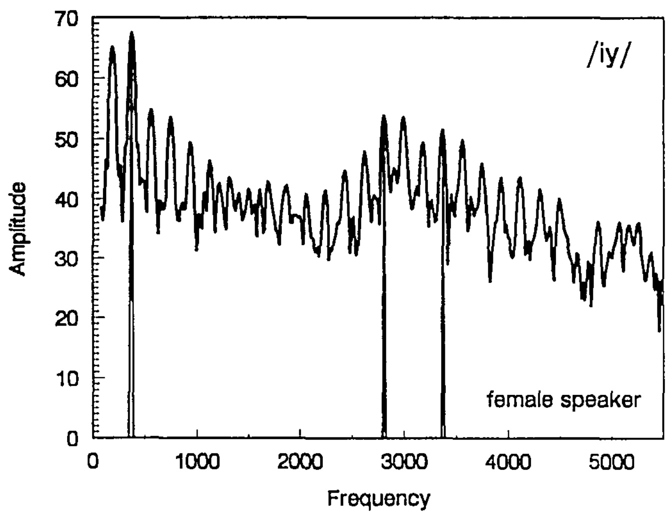

Figure 2.5(c) Formants of a female speaker for vowel /iy/.

In 1976, Markel and Gray advanced a well known speech model called linear prediction of speech. In that model the vocal tract is mathematically described by a linear transfer function (2-1)

$$H(z) = \frac{1}{1+\sum_{i=1}^{p} a_i z^{-1}} \tag{2.1}$$

where a_i's, i = 1 ... p, are the predictor coefficients and p is the order of the model. Since formants are the resonant frequencies of the vocal tract, each formant corresponds to a complex-pole pair of equation (2-1). Therefore, formants can be computed from the linear prediction model. There are two methods to compute the formants. One is called peak picking. That is, the magnitude of the frequency response is computed using the transfer function of equation (2-1); the formants are the frequency locations of the peaks of the smoothed spectrum $|H(e^{jw})|$. The second method, called root solving, is to compute the complex-pole pairs of equation (2-1), i.e., to find the roots of the polynomial A(z), and then to identify formants as the resonant frequencies of these complex-pole pairs. Each of these two methods has some advantages and disadvantages in estimating formants. The disadvantages of peak picking are closely spaced formants often appear as one peak in $|H(e^{jw})|$ and spurious peaks in the spectrum may also be erroneously selected as formants. The disadvantage of the root solving method is that an extraneous complex pole pair can be misidentified as a formant. However, these spurious peaks or extraneous pole-pairs can be excluded in formant estimation using some of the well defined characteristics of formants, such as narrow bandwidth, large amplitudes, and continuity over time. Since roots of the LP model spectrum contain some extraneous complex pole-pairs from the transfer function, these roots are called raw formants or formant candidates.

In the work reported in this thesis, the root-solving method was used to compute formants, since this method has been shown to give superior results (Zahorian and Jagharghi, 1993). The roots of the LP polynomial were computed to obtain up to 5 formant candidates per frame. Table 4.2 lists 5 formant candidates of 10 vowels for an adult male, and a adult female and an 8-year old child.

2.2.2 DCTCs

DCTCs are the Discrete Cosine Transfer Coefficients and represent the smoothed spectral shape of the speech signal or global shape of the spectrum. Pols (1977) used a principal-components spectral shape representation of vowel spectra. The principal-components data were first scaled and rotated to best match the formant data. Zahorian and Gordy (1983) showed that a cosine basis vector representation of the spectrum is nearly identical to a principal-components representation. DCT coefficients are the discrete cosine transform of a selected segment of the spectrum represented by cosine basis vectors. Zahorian and Jagharghi (1990, 1992, 1993) have shown that vowel classification based on DCTCs results in about 2 to 4 percent higher rates versus classification based on formants. Beck also pointed out the DCTC method is relatively convenient for real-time applications (Beck,1992).

DCTCs are also equivalent to low-order cepstral coefficients. The cepstrum is defined as Fourier transform of the logarithm of the magnitude of the spectrum (Oppenheim, 1989). To approximate the psychophysical properties of the ear's response to sinusoids, nonlinear scaling of both frequency and amplitude axes are applied. Therefore, the cosine expansion used in the DCTC computations was applied to the amplitude-scaled and frequency-warped magnitude spectrum of a Hamming-windowed frame of the speech signal. The speech signal was also pre-emphasized at higher frequencies, using the transfer function ($1 - .95z^{-1}$), again to approximate the ear sensitivity. The length of window used was 40 ms. A 1024 point FFT was used to compute the spectrum of the windowed speech signal. The FFT output was converted to a log amplitude scaled spectrum as mentioned above. Nonlinear frequency scaling was accomplished using frequency warping. Two warping methods were used -- one based on a Bark frequency scale and the other based on the bilinear frequency transformation. Bark scale warping, long used in the speech science community (Zwicker, 1961; Syrdal and Gopal, 1986), models the frequency resolution of the ear. The relation between Bark frequency and frequency in kHz is given by the following equation:

$$B=13\tan^{-1}(0.76f)+3.5\tan^{-1}(\frac{f}{7.5})^2 \qquad (2.2)$$

Bilinear frequency warping, the other warping function used in some of the experiments reported in this thesis, is more flexible with regard to the degree of warping and is specified by the formula

$$f'=f+\frac{1}{\pi}\tan^{-1}(\frac{\alpha Sin2\pi f}{1-\alpha Cos2\pi f}) \qquad (2.3)$$

where f is the original normalized frequency, f' is the warped normalized frequency, and α is the control of degree of warping. In most of our experiments, $\alpha = 0.45$ was used, which resulted in the best DCTCs for computing global spectral shape. Note, however, that Bark frequency warping is most similar to bilinear warping if $\alpha = 0.55$.

The definition of DCTCs is given by the equation

$$H'(f')=\sum_{n=1}^{N} a_n Cos(n-1)\pi f' \qquad (2.4)$$

where H'(f') is the magnitude spectrum of the nonlinear warped spectrum. N is the number of DCTCs to be computed for each frame of speech.

2.2.3 Pitch

Voiced sounds, such as vowels are nearly periodic in the time domain. The fundamental frequency, F0, is an important parameter needed to represent the signal. F0 is usually referred to a pitch. (Although, as mentioned previously, this is not rigorously correct.) In our work, pitch was computed from a form of the SIFT fundamental frequency algorithm (Markel, 1972). That is, the LP residual was computed for a window of speech (50 ms for male, 40 ms for female and child) in the steady-state portion of each vowel with a 12th-order LP inverse filter. The details of the algorithm for computing F0 were developed and investigated by Zahorian and Gordy, 1983; and Effer, 1985.

2.3 Data Recording

The data used in this study was collected from one adult male, one adult female, and one child speaker. Recordings were made in a sound-treated room. Each speaker was asked to hold each vowel sound "steady" for at least one second in response to a computer prompt. The sampling rate was set at 32 kHz with a 12 bit A/D, digitally lowpass filtered at 7.5 kHz, and decimated to a sampling rate of 16 kHz. Each speaker produced 10 vowels for an overall total of 30 tokens. The data were stored in binary form with 128 words of header information for each token.

2.4 Format of Experiments

The purpose of the experiments in this study was to evaluate the features mentioned above via two types of listening tests and an automatic identification experiment. The listening experiments were performed in the same sound-treated room used to make the recordings. The listener heard the sound through earphones (Poineer, SE-405) and entered a response via a keyboard entry. The tested vowels were randomized separately for each listener, in order to eliminate possible biases due to order effects. One listening test was called forced choice. After the listener heard a randomized vowel sound, he or she was "forced" to identify the sound from a closed set of possibilities shown on the computer screen. The listener was allowed to listen to each sound as many times as desired, but was encouraged to listen only once. Figure 2.6 shows the screen prompt given.

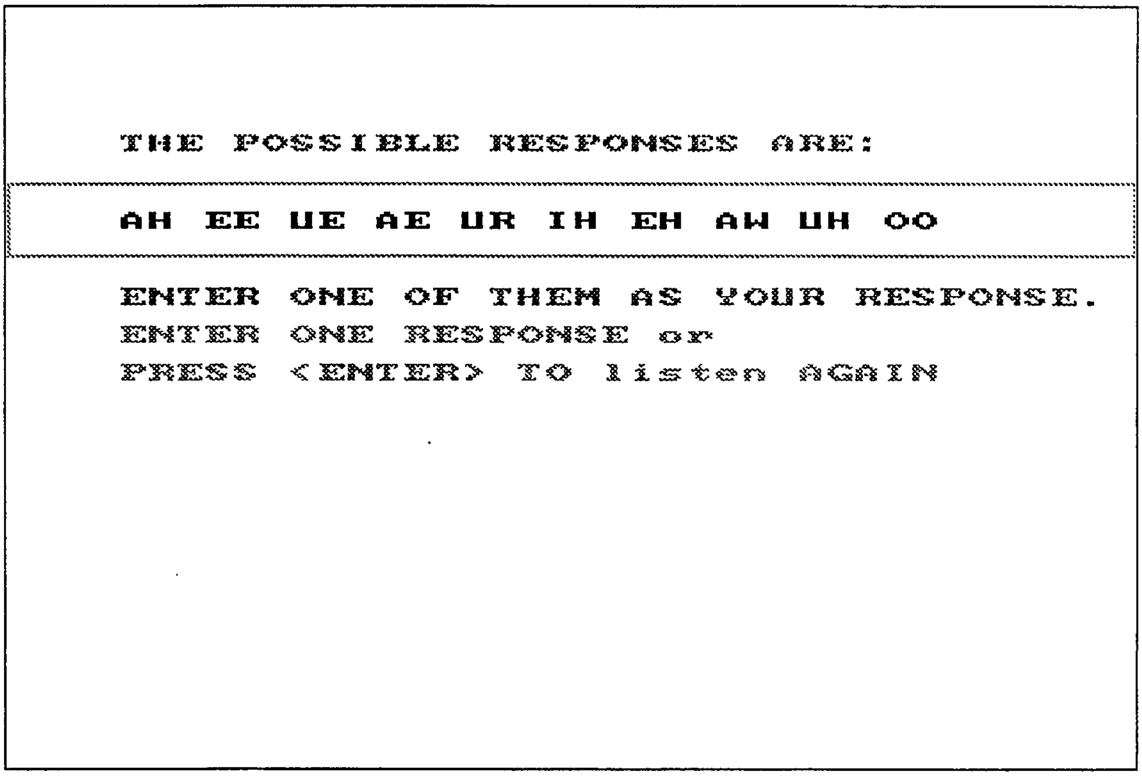

Figure 2.6 The listening experiment environment screen.

Another type of listening test was called the AXB test. This type of test was used to compare two different synthesis conditions and to determine which of these resulted in speech more similar to the original unmodified speech. In this experiment, the listener hears three sounds sequentially-- "A", "X", and "B". The listener must respond as to whether the first ("A") or the last ("B") is more similar to the middle sound ("X"). This type of discrimination test can be used to make finer distinctions than a forced choice test. All groups of 3 tokens were also presented with the role of "A" and "B" interchanged to eliminate biases due to order effects within a group. For example, if a listener cannot distinguish between the three members of a group, the responses may be biased to the "B" choice. The time interval between presentations of sounds was set to 1.5 seconds. Figure 2.7 shows the response screen presented to the listener. A scoring program automatically tabulated the results.

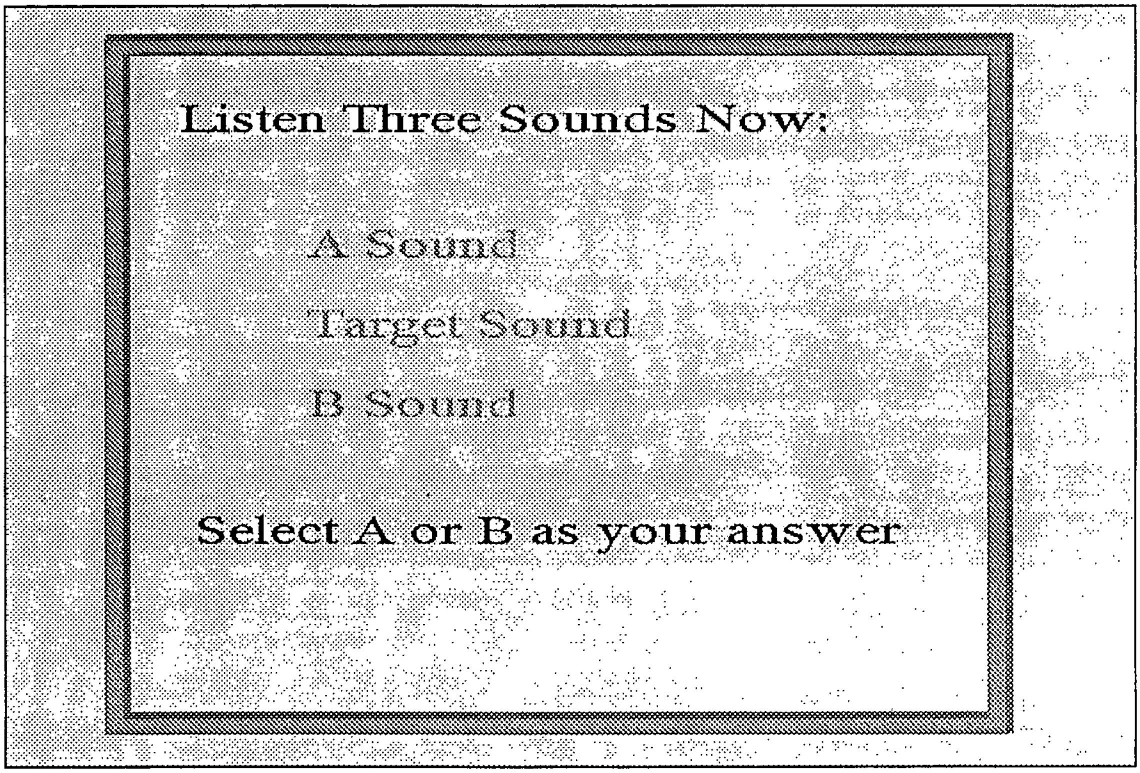

Figure 2.7 AXB experiment response screen.

CHAPTER THREE

Experiments on Formants Versus DCTCs

In previous studies (Nossair and Zahorian, 1991; Zahorian and Jagarghi, 1991, 1992, 1993), it has been determined that automatic classification of vowels and stop consonants is more accurate if global spectral shape features are used rather than formant frequencies. In one pilot study, Jagharghi and Zahorian (1990), also found that if vowels synthesized with conflicting cues to vowel identity in terms of formants and global spectral shape, perception of the tokens more closely follows the spectral shape cues than formant cues. These results thus supported our hypothesis, given in chapter one, that global spectral shape cues are more important to the phonological perception of vowels than are formant frequency cues. This hypothesis was tested using vowel tokens synthesized from a sinusoid model under four conditions to preserve various aspects of overall spectral shape or the first three formant frequencies and amplitudes. The forced choice listening experiment was used to evaluate the intelligibility of original and synthesized tokens. Five vowels /aa, iy, uw, ae, er/ of an adult male and adult female were used in this experiment.

Eleven DCTCs were computed as the first 11 coefficients in the cosine transform of the nonlinearly scaled spectrum over the frequency range of 80 to 4200 Hz for the male speaker and the range of 80 to 5400 Hz for the female speaker. The spectrum recomputed from the DCTCs is thus a smoothed version of the FFT log/Bark spectrum.

Formants used in this experiment were computed based on the method mentioned in 2.2.1. The three computed formant values for each token are given in Table 3.1 for each speaker.

Table 3.1 Formants of five vowels for male and female speakers.

	Male			Female		
	F1	**F2**	**F3**	**F1**	**F2**	**F3**
/aa/	830	1307	2654	1042	1405	3082
/iy/	293	2202	3050	233	2805	3795
/uw/	328	1084	2206	281	1370	2833
/ae/	654	1898	2531	927	2156	2876
/er/	515	1340	1540	423	1408	1699

This chapter describes the test of this hypothesis, strategies for the synthesis conditions, the listening experiments, summarizes the results, and concludes with the implications of these results on the hypothesis for this study.

3.1 Synthesis With Sinusoids Based on Formants and DCTCs

As mentioned in Chapter 2 the speech signal is periodic and can be approximated as a sum of sinusoids. In order to create a periodic time domain signal, the sinusoids were chosen to be integer multiples of the fundamental frequency, that is, harmonic frequencies (1). The minimum phase function of the envelope of the spectrum was originally used for the phase of the sinusoid. This phase is uniquely specified from the magnitude spectrum (Oppenheim and Schafer), and is considered to be a good approximation to the actual phase for speech signals. However, some pilot experiments showed that the minimum phase function was not important for vowel perception. We therefore eliminated the minimum phase function in the rest of our experiments and instead used zero phase. Each stimulus was one second long, including a 25 ms linear on/off ramp. The amplitude of each synthesized token was scaled to match the amplitude of the corresponding original token. There were four synthesis conditions in this experiment, which are described in the following sections.

---

(1) Some experiments were tried without preserving this periodicity and the resultant speech was of very poor quality.

3.1.1 Uniformly Spaced Sinusoids

In this case every harmonic of the fundamental was used over the frequency ranges mentioned. In particular for the male speaker (F0 = 110 Hz), 37 harmonics spanning the frequency range of 110 to 4200 Hz were used; for the female speaker (F0 = 155 Hz) 33 harmonics spanning the frequency range of 155 Hz to 5400 Hz were used. The sinusoidal harmonics were adjusted in amplitude to match the smoothed DCTC spectra, as illustrated in figure 3.1. The harmonics are also a good match to the spectral envelope of the FFT spectrum, except for a lower amplitude. Figure 3.2 depicts the FFT spectrum of the synthesized speech as well as the DCTC spectrum of the original and synthesized speech. Note that although the detailed spectrum of the synthesized vowel is considerably different than the original high-resolution spectrum, both the DCTC spectrum and envelope of the spectrum of the synthesized vowel are quite similar to the corresponding spectra of the original token. Thus both the envelope and global spectral shape of the original token are preserved. However, due to the large degree of smoothing in the DCTC spectrum, formant peaks are not as well-preserved.

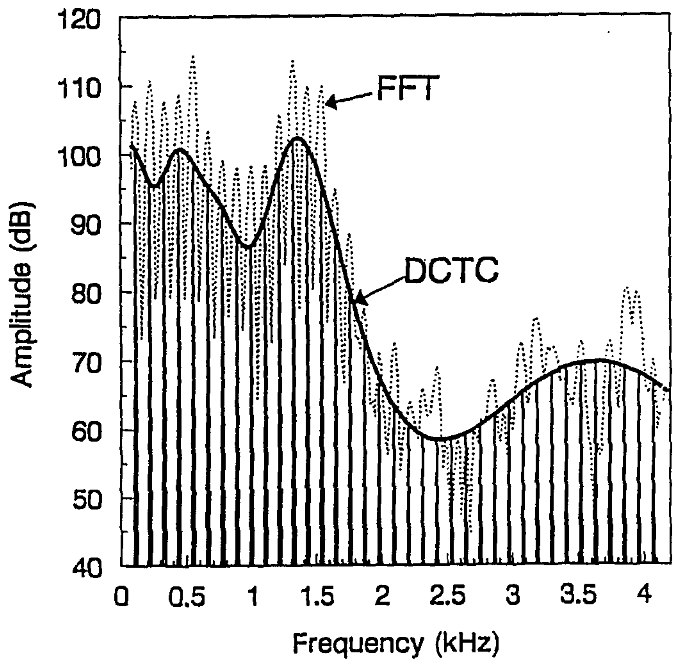

Figure 3.1 Illustration of frequencies and amplitudes of uniformly spaced sinusoids.

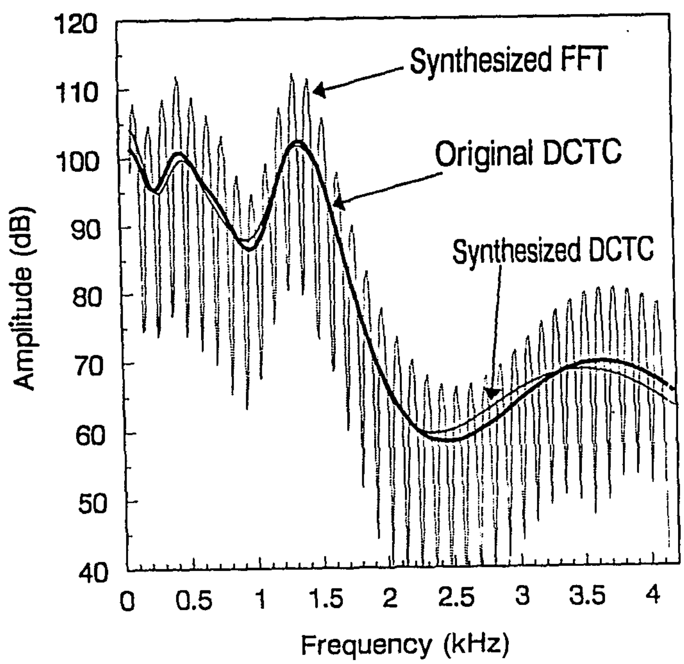

Figure 3.2 DCTC spectrum for original and synthesized speech using uniformly spaced sinusoids.

3.1.2 Bark Spaced Sinusoids to Preserve Spectral Envelope

Our objective in the second case was to replicate case 1 above, but with far fewer harmonics. In particular, we wished to equally space sinusoids on a Bark scale, thus approximating the frequency resolution of the ear versus frequency. Presumably sinusoidal components could be spaced much farther apart at high frequencies than low frequencies with no loss in intelligibility. However, equal spacing on a Bark scale could not be used with the constraint of preserving the harmonic structure of the signal, as required for synthesizing high-quality speech. To roughly approximate a Bark spacing and still preserve the harmonic structure, sinusoids were spaced one harmonic apart for low frequencies, two harmonics apart for middle frequencies, and three harmonics apart at high frequencies. For both speakers a total of 16 sinusoids were used. Figures 3.3 and 3.4 depict the spectral plots for this case. Note from figure 3.4, that although the envelope of the spectrum is preserved quite well, the global spectral shape spectrum is considerably altered from the original.

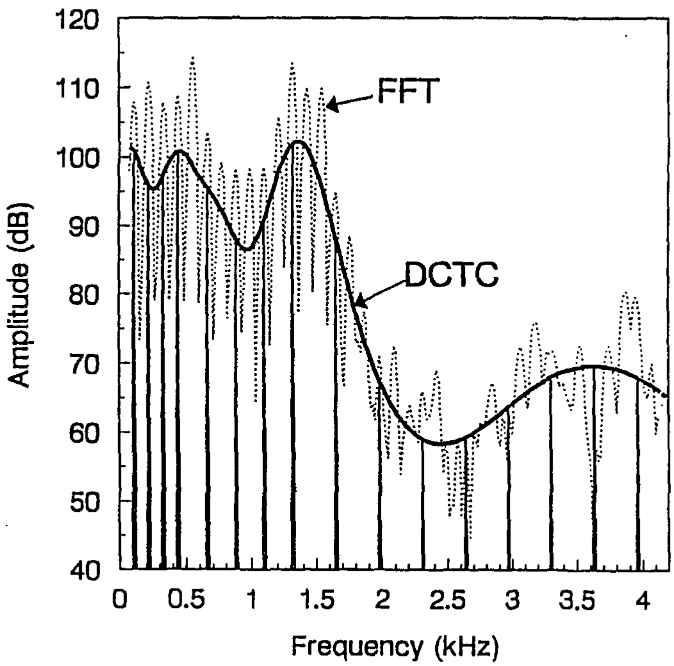

Figure 3.3 Illustration of frequencies and amplitudes of Bark spaced sinusoids which preserve spectral envelope.

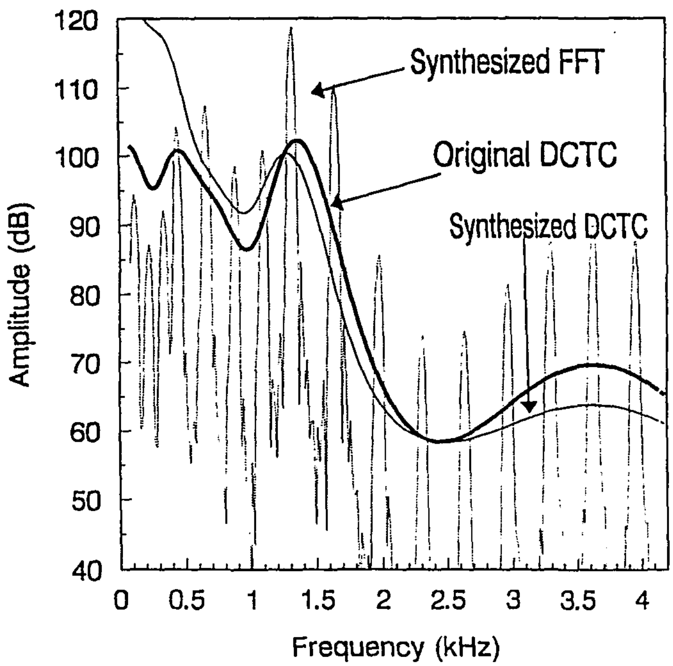

Figure 3.4 DCTC spectrum for original and synthesized speech of Bark spaced sinusoids which preserve the spectral envelope.

3.1.3 Bark Spaced Sinusoids to Preserve Global Spectral Shape

This case is a repeat of case 2 described above, except the amplitudes of the sinusoidal components were adjusted to preserve the spectrum computed from the DCTCs of the synthesized speech. In particular the amplitudes of the lower frequency tones were reduced and the amplitudes of the higher frequency tones were increased to compensate for the nonuniform spacing of tones. Figure 3.5 and figure 3.6 depict the spectral plots for this case. Note that the DCTC spectrum of the synthesized speech is a good match to the DCTC spectrum of the original speech, but the envelope of the spectrum of the synthesized speech is quite distorted relative to the original speech.

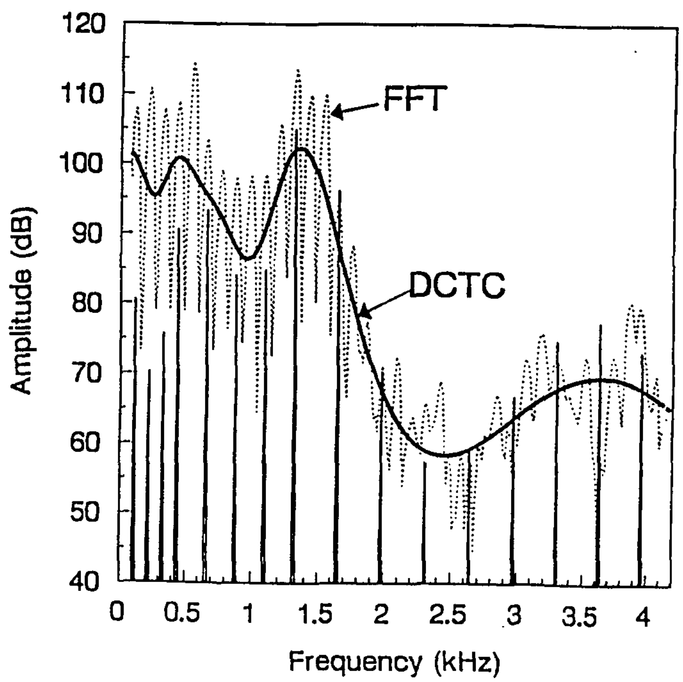

Figure 3.5 Frequencies and amplitudes of Bark spaced sinusoids which preserve global spectral shape.

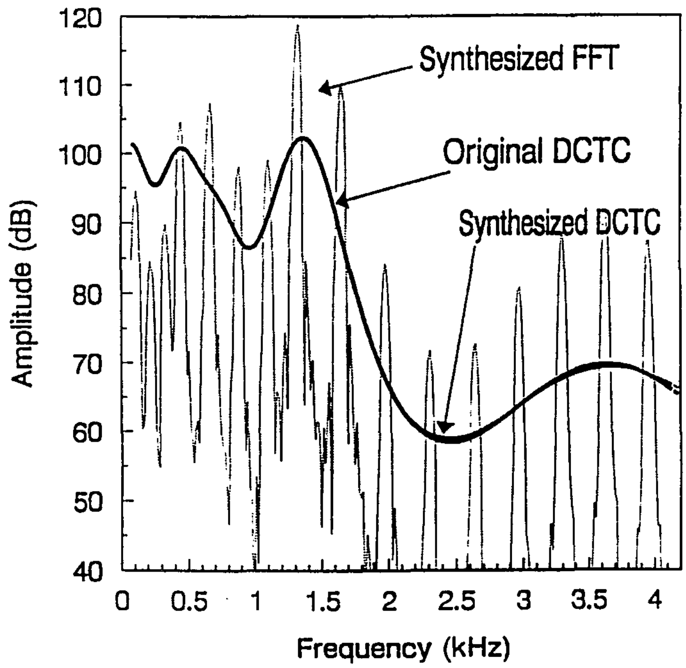

Figure 3.6 DCTC spectrum for original and synthesized speech of Bark spaced sinusoids which preserve global spectral shape.

3.1.4 Sinusoids to Preserve Formant Frequencies

For this case, three tones were used to synthesize each token. The frequencies were chosen to match the formants of the original token, except as adjusted to be a multiple of the fundamental. The amplitudes were chosen to match the DCTC spectrum at the formant frequencies. The spectrum of the synthesized speech preserves the formant peaks. Figure 3.7 and figure 3.8 depict the relevant spectral plots. Note that the only dominant features preserved in the spectrum are the formants. Both the envelope and global spectral shape are distorted relative to the original speech.

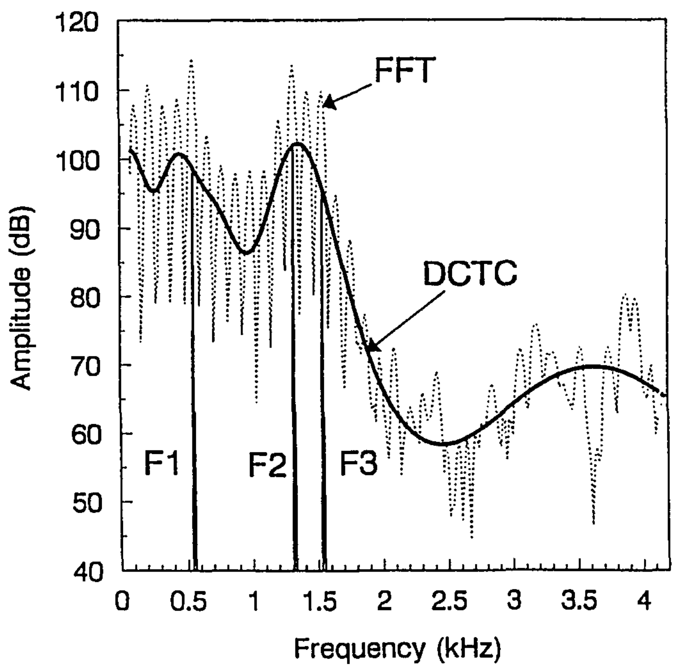

Figure 3.7 Frequencies and amplitudes of sinusoids which preserve formant frequencies.

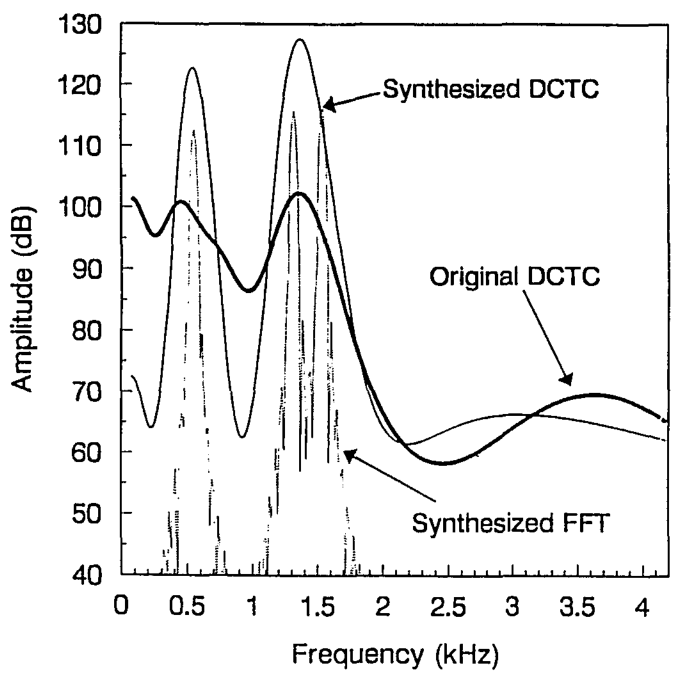

Figure 3.8 DCTC spectrum of original and synthesized speech from sinusoids which preserve formant frequencies.

3.2 Listening Experiment

A total of 25 tokens were created from each speaker (5 vowels each with 4 synthesis conditions plus the original token). Each token was replicated once to produce 50 stimuli per speaker. These stimuli were randomized within a block of 50 for each speaker. Six listeners, all previously used in other similar experiments, were used as subjects. A short training session was used wherein the listener listened to each block of 50 stimuli in a forced choice paradigm. After each token was presented, the listeners entered a 2 character response code and indicated readiness for the next stimulus via a keyboard command. To eliminate order effects, a separate randomization was used for each block of 50 and for each listener. The listeners always heard the male speaker block before the female speaker block.

3.3 Results

The results of the listening experiment were scored by a scoring program which computes the recognition rate and confusion matrices. Scoring was based on the four synthesis conditions and the original speech. It also scored each speaker individually and the average for all speakers. The results of the listening experiment as percentage "correct" are given in figure 3.9 in barograph form. Confusion matrices, averaged over both speakers and all listeners are given in Table 3.2 - 3.6 for the various synthesis conditions. The recognition rates range

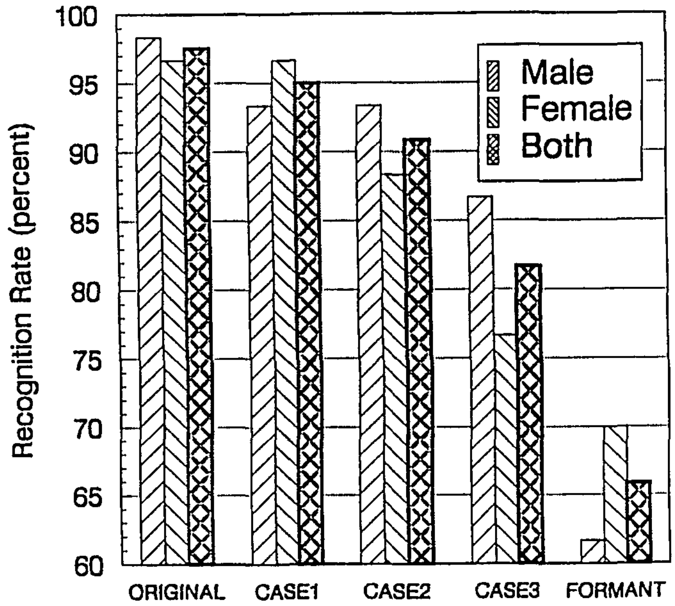

Figure 3.9 Experimental results

from 97.5% for the original speech to 65.8% for speech synthesized from 3 sinusoids which match the formant frequencies. Speech synthesized from uniformly spaced sinusoids, resulting in a match of both the envelope and global spectral shape of the original speech, is nearly as intelligible as the original speech. Speech synthesized from 16 nonuniformly spaced sinusoids is less intelligible than speech synthesized from the larger number of uniformly spaced sinusoids. However the speech from 16 sinusoids which match global spectral shape is considerably less intelligible than the speech from 16 sinusoids which matches the spectral envelope. The speech synthesized from the formant sinusoids, which results in considerable distortions of both the spectral envelope and global spectral shape, is of lowest intelligibility.

Table 3.2 Confusion matrix for original tokens

	/aa/	/iy/	/uw/	/ae/	/er/
/aa/	100.0				
/iy/		100.0			
/uw/			100.0		
/ae/	12.5			87.5	
/er/					100.0

Table 3.3 Confusion matrix for case 1, tokens synthesized with 30+ uniformly spaced sinusoids to match both spectral envelope and global spectral shape.

	/aa/	/iy/	/uw/	/ae/	/er/
/aa/	100.0				
/iy/		100.0			
/uw/			83.3	4.2	12.5
/ae/	8.3			91.7	
/er/					100.0

Table 3.4 Confusion matrix for case 2, tokens synthesized with 16 nonuniformly spaced sinusoids to match spectral envelope.

	/aa/	/iy/	/uw/	/ae/	/er/
/aa/	95.8				4.2
/iy/		100.0			
/uw/			79.2		20.8
/ae/	8.3			87.5	4.2
/er/			8.3		91.7

Table 3.5 Confusion matrix for case 3, tokens synthesized with 16 nonuniformly spaced sinusoids to match global spectral shape.

	/aa/	/iy/	/uw/	/ae/	/er/
/aa/	91.7			4.2	4.2
/iy/	4.2	91.7	4.2		
/uw/	12.5	12.5	45.8	20.8	8.3
/ae/	4.2		8.3	83.3	4.2
/er/	4.2				95.8

Table 3.6 Confusion matrix for case 4, tokens synthesized with 3 nonuniformly spaced sinusoids to match formant frequencies and amplitudes.

	/aa/	/iy/	/uw/	/ae/	/er/
/aa/	58.3	4.2	4.2		33.3
/iy/		95.8	4.2		
/uw/			95.8		4.2
/ae/	4.2	20.8	16.7	4.2	54.2
/er/			25.0		75.0

3.4 Conclusions from this Experiment

Our original hypothesis that vowel perception is closely linked to global spectral shape was only partially supported by the experiment. The experimental results do indicate that preservation of vowel formants alone is not sufficient to reliably cue vowel identity. However, preservation of global spectral shape is also not sufficient to reliably cue vowel identity (case 3). Of our experimental conditions, vowel intelligibility remained high only if both spectral shape and the spectral envelope were preserved. The results indicate that many aspects of the spectrum must be preserved to retain high vowel intelligibility, thus favoring a more "complete" spectral description than is given simply by specifying 3 formant frequencies. However, the method used to measure this global spectral description must be modified from our current DCTC method, if the underlying spectral features are to be closely correlated with perception.

From the experiment we concluded that neither formants nor global spectral shape were sufficient for vowel identity and that both spectral shape and spectral envelope are required. Therefore, as described in the next chapter, we developed and investigated a new algorithm to compute the DCTCs, in the quest for a set of global spectral shape features which are more correlated with perception.

CHAPTER FOUR

Refining Acoustic Correlates for Vowel Perception

The basic conclusion from the last chapter was that neither formants nor global spectral shape features are sufficient cues to predict vowel perception. Therefore, we needed to further develop a formulation of a feature set to predict vowel perception. In this chapter we investigate a new algorithm to compute the DCTCs which we call the DCTC peak algorithm. That is, we use only peaks of the spectrum to obtain the DCTCs. Therefore, rebuilding a smoothed spectrum from these DCTCs matches the envelope of the original spectrum. In the process of developing this new algorithm, we have tested several additional criteria for selecting the amplitudes and frequencies of sinusoids which are required to synthesize intelligible vowel sounds.

4.1 DCTC Peak Algorithm

In chapter 3 we have shown several figures which plot FFT spectra and spectra derived from DCTC spectral shape coefficients. From these figures we see the DCTC spectra track the FFT spectra smoothly, but the peaks of the FFT spectra are not well tracked. We did multi-tone vowel synthesis such that the amplitudes of the synthesis components were adjusted so as to preserve the DCTCs of the spectrum of the synthesized speech rather than the amplitudes of the original harmonic frequencies. The resultant synthesized speech did not preserve vowel intelligibility to a high degree. Speech intelligibility was much higher if the amplitudes were adjusted so as to match the original harmonic amplitudes and frequencies. These experiments did show that the amplitudes, as well as frequencies, of the harmonics play an important role in the multi-tone sinusoidal synthesis model. We hypothesized that DCTCs which encode the harmonic peaks of vowel spectra will be good cues for multi-tone vowels. We therefore call these DCTC peaks.

In this chapter we first present the algorithm for computing the DCTCs with this approach. The algorithm is formulated mathematically as follows. Consider a set of M orthonormal basis vectors over [0, N-1] denoted by

$$\Phi_k(j)$$

$$\text{where} \qquad 0 \le j \le N-1$$

$$0 \le k \le M-1.$$

Note that these basis vectors need not be cosines, although in our experimental work, the basis vectors were cosines as used previously. The goal of the DCTC peaks algorithm is to perform a minimum mean square error fit of the these basis functions to the spectral peaks. Therefore, for a spectral frame, consisting of N log amplitude samples, a set of NL harmonic peaks must be determined.

Let S[i], $0 \le i \le NL-1$, be the indices of the peaks,

A[S[i]] be the amplitudes of the peaks,

W[i] be a weighing function.

Finally compute coefficients c(k), $0 \le k \le M-1$, such that

$$E=\sum_{i=0}^{NL-1} W[i]\{A[S[i]]-\sum_{k=0}^{M-1} c(k)\,\Phi_k[S[i]]\}^2$$

is minimized.

The modified DCTC coefficients, c(k), are computed solving the matrix equation

AX = b, where

$$A_{jk}=\sum_{i=0}^{NL-1} W[i]\,\varphi_j[S[i]]\,\Phi_k[S[i]],$$

$$X_k = c[k],$$

$$b_k=\sum_{i=0}^{Nl-1} W[i]\,A[S[i]]\,\Phi_k[S[i]],$$

$$0 \le j \le M-1,\ 0 \le k \le M-1.$$

Figure 4.1 depicts the original DCTC spectrum (as computed from DCTCs used in the work described in the last chapter) and the spectrum recomputed from the DCTC peaks algorithm. The figure clearly shows that the DCTC peaks spectrum is a much better match to the envelope of original spectrum than is the spectrum computed from the "regular" DCTCs. Therefore, we hypothesize that DCTCs computed with the DCTC peaks algorithm will be good acoustic cues for vowels, since our previous results imply that the spectral envelope should be preserved for good vowel intelligibility. In the next section we discuss a variety of synthesis conditions to test this hypothesis.

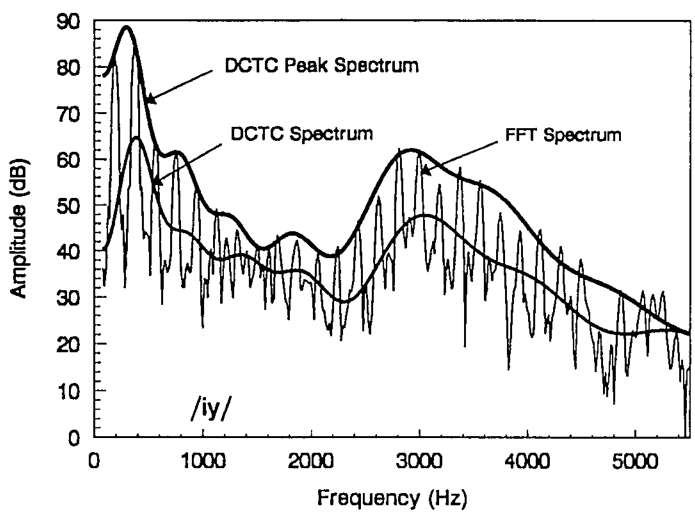

Figure 4.1 Comparison of normal DCTCs spectrum and peak DCTCs spectrum.

An essential component of the above algorithm is the procedure for finding the peaks of the FFT spectrum. These peaks, for voiced speech, are harmonically related. However, for real signals, the peaks are not necessarily at exact multiples of some fundamental frequency. Therefore we used the following procedure to determine peaks. First the fundamental frequency F0 was estimated using the pitch detection routine previously mentioned. Then we searched for a spectral peak in the range from F0/2 to 3F0/2. The frequency of this peak was considered to be the location of the first peak, P1. The next peak was determined by searching in a range of (P1 + F0) ± F0/2. This procedure was iterated until the entire spectrum was searched. This algorithm was graphically inspected for numerous cases, and found to be robust for locating peaks in the spectral envelope, even if F0 was in error.

4.2 Synthesis Experiment

This experiment used ten vowels, /aa, iy, uw, ae, er, ih, eh, ao, ah, uh/, each spoken by one adult male, one adult female, and one child (eight years old), for a total of thirty tokens for each synthesis condition. The spectrum of each token was computed, using a 40 ms window centered in the token, over the frequency range of 80 to 4200 Hz for male and 80 to 5500 Hz for both the female and the child.

The synthesis conditions tested were based on five different groups for a total of 19 specific cases. That is, each group contained several specific cases. In the first group, the control group, there were two cases. One was the original speech, and another was repeating one period of the vowel. The second group consisted of tokens synthesized with varying numbers of harmonic peaks. The amplitude of the peaks were based on two categories -- the original spectral amplitudes and DCTC peak spectral amplitudes. The number of peaks varied from 5 to 16. The third group consisted of tokens synthesized from formant harmonics. In this group we also used tokens with either one or two harmonic peaks, adjacent to each formant, added. The fourth group, with only one condition, was called Bark spaced harmonics. That is, we divided the full frequency range into intervals equally spaced on a Bark scale. The last group consisted of some additional combinations of the few largest harmonic peaks. Table 4.1 summarizes these groups and the number of cases in each group. All synthesized speech had the same length and same on/off ramp and was scaled to match the dynamic range of the A-to-D converter ± 5(volts). In the next section, we give more details of these test conditions.

4.2.1 Original Vowel and Repetition

In this group, the original speech has been modified. The length of each stimulus vowel was changed from 1 second to 560 ms. These segments were taken from the center of the original plus a linear on/off ramp of 25 ms on each side. The second case of the first group consisted of tokens formed by repeating one period of the vowel waveform, taken from the center, enough times to form a segment of the same length as for the first condition.

Table 4.1 Summary of synthesis conditions

Group	Description of Group	Cases
1.	Original Vowel and Repetition	#1, #2
2.	Varying Numbers of Harmonic Peaks	#3, #4, ... ,#12, #13
3.	Formant Harmonics	#14, #15, #16
4.	Bark Spaced Harmonics	#17
5.	Largest Peaks	#18, #19

Case Number	Description of Case
1.	Original vowel
2.	One period of original vowel repeated
3.	All original vowel FFT peaks
4.	5 Largest original vowel FFT peaks
5.	6 Largest original vowel FFT peaks
6.	7 Largest original vowel FFT peaks
7.	8 Largest original vowel FFT peaks
8.	9 Largest original vowel FFT peaks
9.	10 Largest original vowel FFT peaks
10.	16 Largest original vowel FFT peaks
11.	All DCTC smoothed spectral peaks
12.	8 DCTC smoothed spectral peaks
13.	16 DCTC smoothed spectral peaks
14.	All possible formant harmonics (3 to 5)
15.	All possible formant harmonics + one peak at each side *
16.	All possible formant harmonics + two peaks at each side *
17.	16 Peaks selected from equal Bark spacing
18.	8 Peaks without 4 largest peaks
19.	4 largest peaks + one peak at each side *

* For some vowels these added peaks are duplicated at some frequencies.

4.2.2 Varying Numbers of Harmonic Peaks

For this group, every harmonic of the fundamental was computed over the frequency range mentioned above for each token. Harmonic peaks were located by searching a small range around the "expected" frequency for each peak, as discussed previously. In particular for the male speaker the fundamental frequency was about 100 to 110 Hz, for the female about 155 to 185 Hz, and for the child about 180 to 200 Hz. The actual fundamental frequency was computed using the pitch estimation algorithm described in chapter 2, section 2.3.

The magnitudes of the peaks were based on two different spectra, the original FFT spectra and DCTC peak spectra. This second method for selecting the amplitudes was mainly a check of the DCTC peak algorithm, since these amplitudes should have been very similar to those of the original FFT spectra. For the DCTC peak method, we used all FFT harmonic peaks to compute the DCTC peak coefficients, and recomputed the spectral envelope from these coefficients. The DCTC peak spectrum preserves both the envelope and global spectral shape. Therefore, it also preserves the formant peaks to a larger extent than does the normal DCTC spectrum. For these cases we selected varying numbers of the largest harmonic peaks and used the corresponding amplitudes and frequencies to control the sinusoidal synthesizer. Pilot experiments indicated that the best strategy for synthesizing intelligible vowels from a limited number of sinusoidal components was to use components corresponding to the largest peaks in the original spectrum. Figure 4.2 shows the largest 8 harmonic peaks selected from the DCTC peak spectrum.

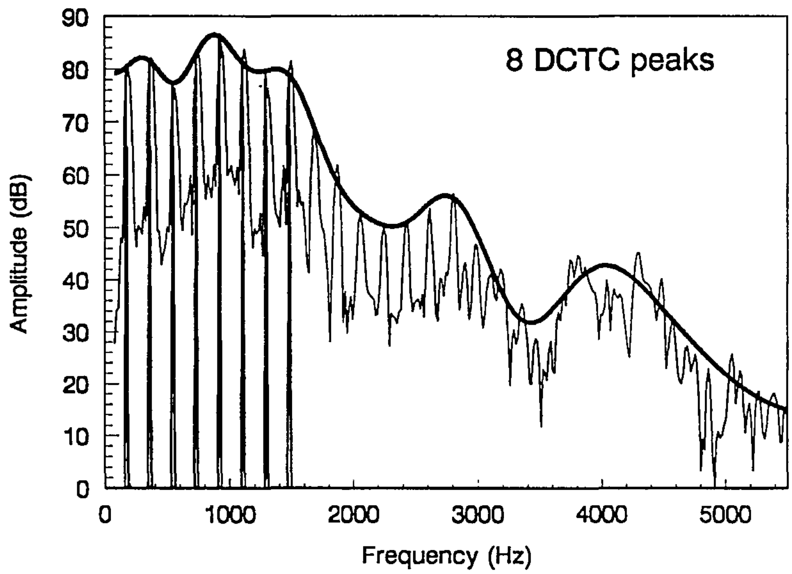

Figure 4.2 Illustration of 8 largest DCTCs spectral peaks for /aa/.

4.2.3 Formant Harmonics

For this group, harmonics at the formant frequencies were used to synthesize each token. The formants were computed as discussed in Chapter 3. For this experiment all possible formant candidates were used. Table 4.2 gives these formant candidates for each token for each speaker. There were three cases in this group. The first was to use only the formant peaks for synthesis. For the second case, two harmonic peaks, one immediately adjacent to each formant on each side, were added to each formant peak. The third case was similar to the second case, except two peaks on each side of each formant were added. Note that the maximum number of peaks for case 2 was 15, whereas the maximum for case 3 was 25. Generally, however, the number of peaks used was fewer, both due to less than five formant candidates and due to the fact the peaks to be added might have resulted from closely spaced formants. Figure 4.3 depicts the relevant spectral plots.

Table 4.2 All possible formants values for 10 American English vowels and three speaker groups.

		/aa/	/iy/	/uw/	/ae/	/er/	/ih/	/eh/	/ao/	/ah/	/uh/
	F1	778	361	398	708	561	533	604	693	686	583
	F2	1251	2173	1073	1713	1297	1771	1705	936	1185	1251
	F3	2259	2992	2038	2521	1548	2550	2634	2571	2403	2237
	F4	2981	3387	3068	2985	3014	3195	3042	3013	3016	3120
	F1	788	363	378	639	516	425	711	605	656	544
	F2	1004	2885	1041	913	1533	526	794	953	814	1226
	F3	1538	3424	1156	2109	2080	2495	2059	1286	1369	2918
	F4	2805		2917	3008	3503	3081	3010	2867	2834	
	F1	1034	473	485	970	719	679	923	954	909	671
	F2	1561	3057	968	1154	1725	1101	1260	1272	1334	1339
	F3	2058	3581	1469	2292	2212	2305	2101	1652	1604	3109
	F4	3595		3222	3454		2560	3419	3472	3491	
	F5			3372			3466				

Note that the peaks were selected from the magnitude of the spectrum and the formant peaks are harmonically related.

4.2.4 Bark Spaced Harmonics

This case is similar to the one described in section 3.1.3 of Chapter 3, except that the amplitudes of the sinusoidal components were adjusted to preserve the spectrum computed from the DCTC peak spectrum. However, these amplitudes also matched the original spectrum peaks very well because the DCTC peak spectrum tracks the envelope of the FFT spectrum. Figure 4.4 depicts the spectral plots for this case.

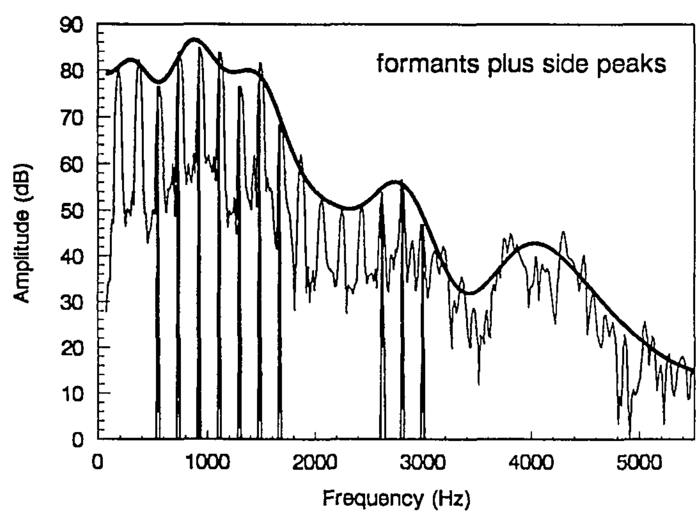

Figure 4.3 Formant frequencies plus side peaks.

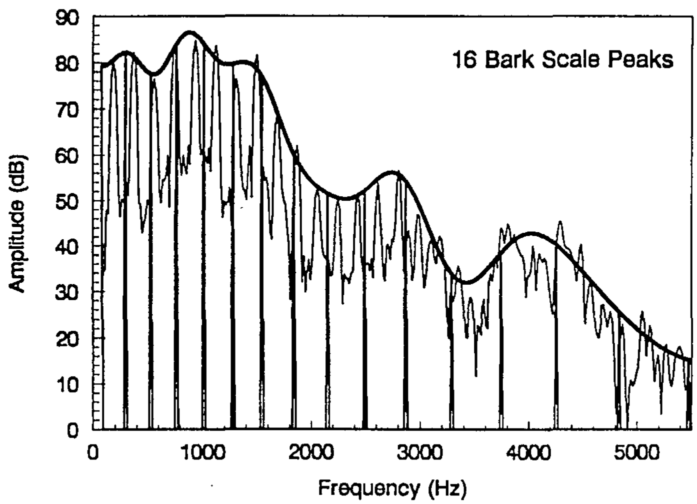

Figure 4.4 Bark scale peaks selected as sinusoids.

4.2.5 The Largest Peaks

This group has two cases that address the issue of the importance of the largest spectral peaks for vowel perception. For one case, the four largest peaks were not used but the next eight largest peaks were used to form the sinusoidal components. For the other case, the four largest peaks and plus one adjacent side peak for each "large" peak were used as the sinusoidal components. Note, that although these four largest peaks contained some of the formant candidates, in general these four peaks were not identical to four formants. That is we simply selected the four largest peaks, without regard to frequency location or spacing or any of the other constraints normally used in identifying formants. In some cases, all four peaks were clustered together and only contained one formant candidate. Figures 4.5 and 4.6 depict the spectral plots and selected peaks for these cases.

4.3 Listening Experiment (II)

A total of 570 tokens were created from each speaker (10 vowels each with 18 synthesis conditions plus the original). These stimuli were randomized within a block of 190 stimuli for each speaker. They were then organized into two equal sub-blocks of 95 stimuli each. Eleven listeners were used as subjects for the forced choice experiment as described in Chapter 3. Each subject took about 15 minutes to complete the experiment. To eliminate order effects, a separated randomization was used for each block of 190 and for each listener.

4.4 Results

The results of the listening experiment were automatically scored to compute the percentage of recognized vowels and confusion matrices. All of the conditions were scored individually for each speaker and also averaged over all speakers. The results of the experiment are summarized in Figure 4.7 and Table 4.3, as to the average recognition rate for each condition. Confusion matrices, averaged over all speakers and all listeners are given in Appendix A for the various synthesis conditions.

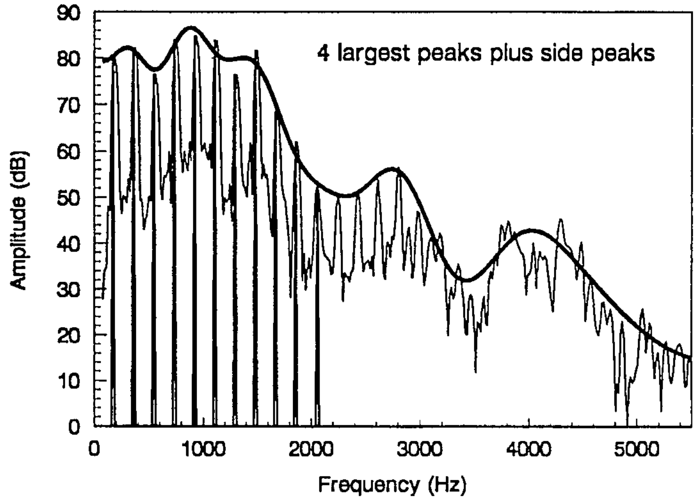

Figure 4.5 Four largest peaks plus side peaks as sinusoids.

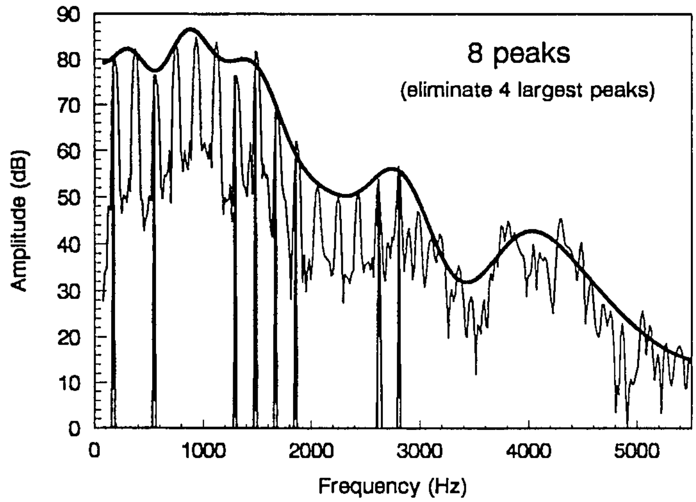

Figure 4.6 Eight peak sinusoidal speech, without the four largest peaks.

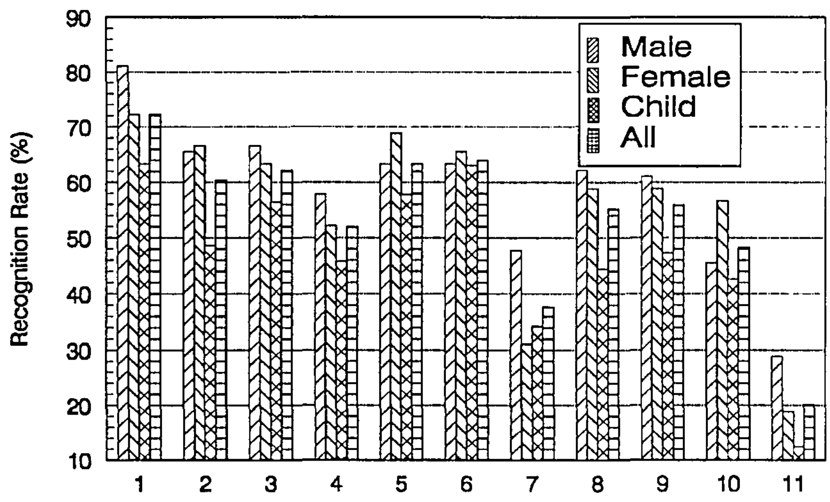

Figure 4.7 Bar graph of experimental results.

From this experiment we can make some conclusions. The perception of synthesis tokens based on DCTC peaks is similar to the perception of tokens synthesized with the original peaks. It implies that the coefficients of the DCTC peaks algorithm can be considered as a new feature set for vowels. The results show that the largest peaks are important in vowel perception. The experiments also show that formants do not supply enough information for vowel perception.

Table 4.3 Results of the experiment

Synthesis conditions	Male	Female	Child	All
Original Speech	81.1	72.2	63.3	72.2
One Period Repetition	65.6	66.7	48.7	60.3
All Original Peaks	66.7	63.4	56.4	62.25
5 Original Peaks	57.8	52.2	45.8	51.9
6 Original Peaks	53.3	57.8	50.2	53.8
7 Original Peaks	60.0	62.2	54.4	58.9
8 Original Peaks	66.7	65.6	53.8	62.0
9 Original Peaks	64.4	61.1	56.9	60.8
10 Original Peaks	64.0	60.5	57.2	60.5
16 Original Peaks	63.4	68.9	57.8	63.3
All DCTC Peaks	61.1	64.5	53.8	63.3
8 DCTC Peaks	67.8	57.8	55.1	60.2
16 DCTC Peaks	63.4	65.6	63.1	64.0
Formants	47.7	32.1	34.2	37.7
Formants Plus Side Peaks	62.2	58.9	44.4	55.2
4 Largest Plus Side Peaks	61.1	58.9	47.3	55.8
16 Bark Scale Peaks	45.6	56.7	42.7	48.3
8 Peaks(w/o 4 Largest Peaks)	28.9	18.9	47.3	31.7

4.5 AXB Experiment

An AXB experiment was used to compare pairs of synthesis conditions and to determine which of these results in speech was more similar to the original unmodified speech. Since this type of experiment can be used to make fine distinctions between two different synthesis tokens, we chose a set of synthesis conditions which gave similar results in the forced choice test such as 5 original peaks versus 8 original peaks, original speech versus one period repetition, etc. Table 4.4 lists these six comparison conditions for the AXB experiment. Each condition had three speakers and each speaker was done separately. Each condition had ten tokens. Each group of three tokens was duplicated once with A and B interchanged. That is, if a token appears in position A first, the second time it must be in position B. Therefore, a total of 120 groups of three were generated for each speaker. Each listener evaluated three speakers. There were four listeners who took part in this experiments. The results are listed in table 4.5.

Table 4.4 List of conditions for AXB experiment.

Case	A	B
1.	Original Speech	One Period Repetition
2.	One Period Repetition	All Original Peaks
3.	Formants Plus Side Peaks	4 Largest Plus Side Peaks
4.	5 Original Peaks	8 Original Peaks
5.	10 Original Peaks	16 Original Peaks
6.	16 Original Peaks	All Original Peaks

Table 4.5 List of results of AXB experiment.

Case	A	A (%)	B (%)	B
1.	Original Speech	71.3	28.7	One Period Repetition
2.	One Period Repetition	56.3	43.7	All Original Peaks
3.	Formants Plus Side Peaks	20.8	79.2	4 Largest Plus Side Peaks
4.	5 Original Peaks	24.6	75.4	8 Original Peaks
5.	10 Original Peaks	40.8	59.2	16 Original Peaks
6.	16 Original Peaks	47.5	52.5	All Original Peaks

From this AXB experiment we can see some differences which were not apparent from the forced choice test. For example, the formants plus side peaks comparison with 4 largest peaks plus side peaks clearly shows that the 4 largest plus side peaks case is preferred. In contrast for the forced choice identification, these two conditions appeared to be about the same. These results also show continuing preference as more and more harmonics are added. Thus, even though the largest peaks are the most important, the other harmonic peaks also improve vowel quality. This result thus implies that a global spectral shape representation, which integrates information from the entire spectrum, is required.

CHAPTER FIVE

Classification

The goal of the experiments described in this chapter was to compare two spectral feature sets using automatic vowel classification. The two feature sets were the normal DCTCs and DCTCs computed so as to encode the spectral envelope. The normal DCTCs are the discrete cosine transform coefficients of the short-time magnitude spectrum of the speech signal. They encode the smoothed magnitude spectrum of the acoustic speech signal. Therefore they preserve the global spectral shape of the speech. However, the multi-tone vowel synthesis reported in Chapter 3 showed that these features are not necessarily good predictors of speech perception. Therefore, as discussed in the previous chapter, we developed and investigated a new methodology to compute the DCTC coefficients which we called the DCTC peak algorithm. Since these DCTCs encode the envelope of the spectrum, and since the envelope spectrum appears to be important for speech perception, as illustrated by the experiments reported in both chapter 3 and chapter 4, these DCTCs are better acoustic correlates of vowel perception than are the normal DCTCs. However, a main reason to determine good acoustic correlates is to improve automatic vowel recognition. Therefore in these experiments, we compared the normal DCTCs and DCTCs computed from the envelope spectrum as features for automatic vowel identification.

In the first section we present an overview of the classification method used in the experiments. Following that, several computation methods for computing the envelope DCTCs are discussed. Some of these methods use the peak DCTC algorithm mentioned in the last chapter and other methods use the normal DCTC computations, but preceded by some preprocessing steps. We compared the two feature sets for four conditions. They are a single frame of clean speech, a single frame of noisy speech with varying signal-to-noise ratio, multi-frames of clean speech, and multi-frames of noisy speech.

5.1 Classifier

The classifier used to evaluate these two DCTC feature sets is called a maximum likelihood classifier. The classification system has two phases: (1) training; and (2) testing. During the training phase the system is presented with the pattern (the DCTCs) for each class from a training set of data. The system computes parameters based on this information. During the testing phase the system uses these parameters to make decisions about the class to which an unknown input pattern is most likely to belong. For the experimental data reported, the test data were from different speakers than those used for training the classifier.

The classifier used in this experiment was already available in the Speech Communication Laboratory at Old Dominion University. The parameters for this classifier are estimated from the DCTC feature vectors of the training set. In the test phase, the classifier assigns unknown patterns to the category with the largest a posteriori probability (i.e. conditioned on observed feature values), according to the multivariate Gaussian model assumed by the classifier, and the model parameters determined in the training phase. James Mike (1985) summarized the implementation of this classifier as follows.

Let C_1, C_2, ..., C_M be M different categories of patterns. Let the feature vector (the DCTC's vector) X be an N-component vector-valued random variable. Let $p(X|C_i)$ be the probability density function of X given the category C_i. Let $p(C_i)$ be the a priori probability of category C_i. Then the a posteriori probability $p(C_i|X)$ can be computed from $p(X|C_i)$ by Bayes rule:

$$p(C_i|X) = \frac{p(X|C_i)}{p(X)} \tag{5.1}$$

where

$$p(X) = \sum_{i=1}^{M} p(X|C_i)\, p(C_i) \tag{5.2}$$

Using Bayes decision rule, the feature vector X will be classified to the category i for which $p(C_i|X) > p(C_j|X)$, for all j not equal to i (5.3)

Equation 5.1 is inserted into Equation 5.3 and the common terms are canceled and taking the logarithm on both sides, we obtain

$$\ln p(X|C_i) + \ln p(C_i) > \ln p(X|C_j) + \ln p(C_j) \qquad (5.4)$$

Assume $p(X|C_i)$ is multi-variate normal; that is,

$$p(X|C_i) = (2\pi)^{-N/2} |R_i| \exp [-0.5(X-X_i)^t R_i^{-1}(X-X_i)^t] \qquad (5.5)$$

where X_i is the mean for category i and R_i is the covariance matrix for category i. Substituting this equation into Equation 5.3 and multiplying by-1 results in

$$\ln |R_i| + (X-X_i)^t R_i^{-1}(X-X_i) - 2\ln p(C_i) < \ln |R_j| + (X-X_j)^t R_j^{-1} - 2\ln p(C_j) \qquad (5.6)$$

The decision based on equation 5.5 is equivalent to saying that the vector X will be assigned to the category i for which the "distance"

$$D_i(X) = (X-X_i)^t R_i^{-1} \quad (X-X_i) + \ln |R| - 2\ln p(C_i) \qquad (5.7)$$

is minimized. This distance, called the maximum likelihood distance, is the criterion by which the maximum likelihood classifier makes its decision. Therefore during the training phase, the training patterns of each category are used to compute centroids and covariance matrices for each category. During the testing phase, the classifier uses the computed centroids and covariance matrixes to compute the distance of the unknown input pattern X to each of the categories. The classifier then assigns the pattern X to that category for which the computed distance is minimum. This classifier is optimum if the conditional probability density functions $p(X|C_i)$ are actually multi-variate Gaussian, as assumed (Duda and Hart, 1973).

5.2 The Computation of Normal DCT and Envelope DCT Coefficients

Both DCTC coefficient sets were computed from the short-time magnitude spectrum of the speech signal. Many variables are involved in the computation of both feature sets. The main variables include: (1) the total number of DCT coefficients; (2) the amplitude scaling method; (3) the frequency warping method; (4) the length of the window; and (5) the frequency range over which the DCTC coefficients are computed. The envelope DCT coefficient computations require an extra variable which controls the selection of the peaks. The total number of DCT coefficients and the method for peak selection were the two main quantities which were varied in this experiment.

5.2.1 The Common Variables

Based on the previous work of Zahorian and Jagharghi (1991, 1993), and pilot experiments conducted in this study, some variables were determined and kept constant for the primary experiments. In particular, these variables were chosen as follows:

(1). The amplitude scaling method was chosen to be logarithmic to approximate the auditory amplitude response.

(2). The frequency warping was bilinear frequency warping with warping coefficient (α) equal 0.45. One experiment was performed to compare $\alpha = .45$ and $\alpha = .30$ in DCTC computations. Figure 5.1 shows some experimental results.

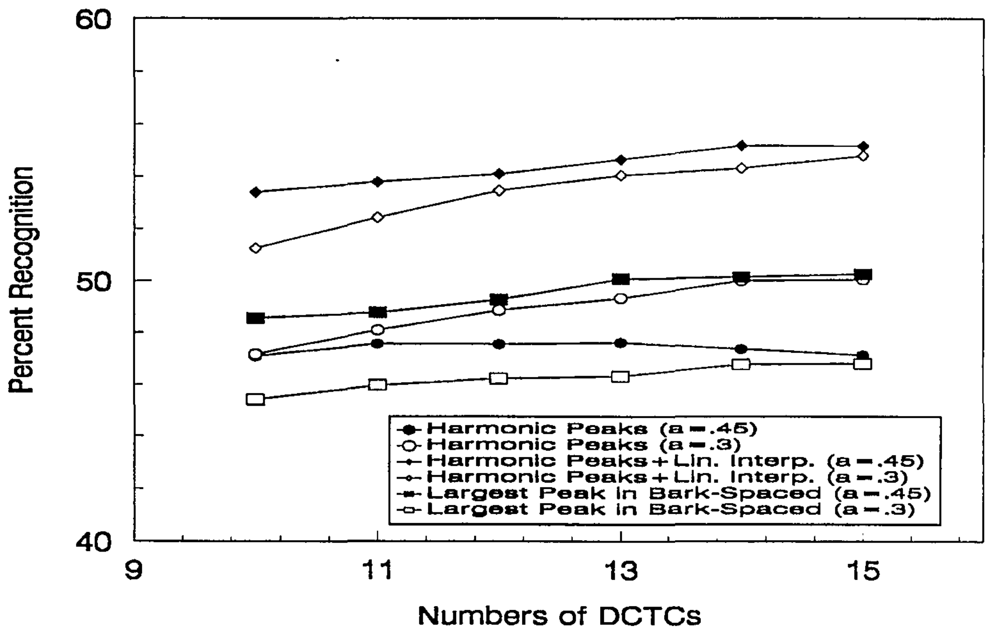

Figure 5.1 Illustration of the effect of the degree of warping on vowel classification(16 vowels) for several envelope DCTC computation methods.

Notice that $\alpha = .45$ is better than $\alpha = .30$ for most of DCTC computation methods. The only situation for which $\alpha = .30$ was preferred was for the harmonic peaks DCTC computations. However, the harmonic peaks method generally was worse than any others for computing the DCTCs, in terms of classification results. Note that the above experiments was based on 16 vowels, whereas all the other experiments reported in this chapter were based on 13 vowels.

(3). The exact frequency range had little effect on results for either normal DCTCs or envelope DCTCs. For normal DCTCs the frequency range was 75 Hz to 6000 Hz. The frequency range for the DCTC peak algorithm was from 0 to 6000 Hz due to a requirement of the procedure.

(4). The length of the window was 30 ms. Usually the length of window was 25 ms for vowel classification experiments performed in our lab. However, since in our experiments we mainly processed a single frame of the speech signal, we added 5 ms more to the signal to slightly improve the frequency resolution. The window started 15 ms before the center of the speech signal and ended 15 ms after the center. For consistency this length also was used in the multi-frame experiment.

These parameter settings were used in both normal DCTC and envelope DCTC computations.

5.2.2 Method for Selecting Peaks

The computation of DCT coefficients was based on the FFT spectrum. For normal DCT coefficients the computation used the entire magnitude spectrum at all points over the selected frequency range. For the 1024 point long FFT used, and the frequency range of 6000 Hz, there were 384 such points in the spectrum. Figure 5.2 depicts the normal DCTC spectrum with its FFT spectrum. In contrast, the envelope DCTC peak computations used only a subset of these 384 points, corresponding to the peaks of the FFT spectrum. There were two basic methods used to select the peaks: (1) harmonically-spaced peaks; (2) Bark-spaced peaks (i.e., peaks equally-spaced on a Bark scale). In addition there were two variations for selecting Bark-spaced peaks. Each of the peak-picking methods had advantages and disadvantages, as discussed below.

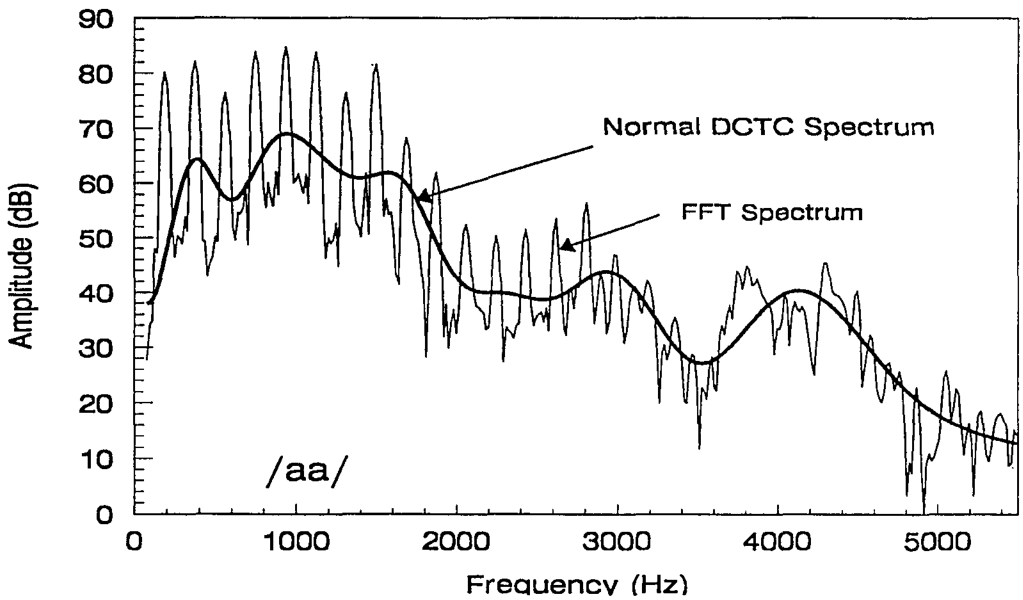

Figure 5.2 Illustration of spectrum from normal DCTCs and original FFT spectrum.

(1). Harmonically-related peaks

For this method, all harmonic peaks of the FFT spectrum were used and the other spectral points were not used. The number of these peaks ranged from less than 30 to about 55, depending on the fundamental frequency F0. These peaks were used both with the DCTC peak algorithm and the normal DCTC algorithm as described in the next section. Recall that the algorithm for selecting these peaks was given in a previous section.

(2). Bark-spaced peaks

This method is based on the frequency selectivity of the human ear. Human ears have more selectivity at low frequencies than high frequencies. The Bark frequency scale approximates this characteristic. Relative to linear frequencies in Hz, equal spacings on the Bark frequency scale are close together at low frequencies and farther apart at higher frequencies. For the Bark-spaced peak method, we divided the entire frequency range into N contiguous, nonoverlaping

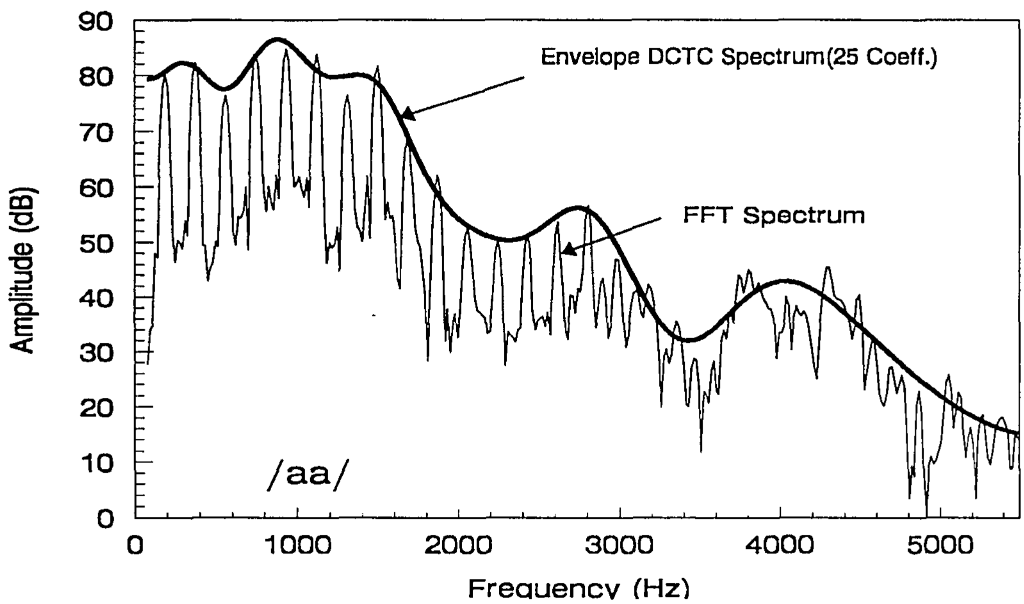

Figure 5.3 Illustration of the envelope DCTCs spectrum with computations based on harmonically related peaks (method 1 in text).

intervals, each of equal width on a Bark scale. Thus, in terms of Hz, the bandwidth of these intervals is relatively narrow for low frequencies and wider at high frequencies. Based on these N Bark-spaced intervals, we selected N peaks using two different methods.

The first one was to choose the largest peak in each interval. These peaks are points on the envelope of the magnitude of the FFT spectrum but are generally not equally spaced in frequency (i.e., not harmonically related).

The second method was to select peak values spaced exactly uniformly on the Bark scale. However, since in general these frequency values did not correspond to harmonic peaks, the spectrum was first preprocessed to locate all harmonic peaks, and then linearly interpolated between these peaks. Thus the peak values were selected from the harmonic peaks/linear envelope of the FFT spectrum.

The advantage of the first Bark-spaced method for selecting peaks is that it is independent of the fundamental frequency F0. It avoids the computation of the fundamental frequency F0 and thus saves computation. The disadvantage is that some harmonic peaks are missed since it only selects one peak for each interval. With the first method, it was also not possible to use too many Bark-spaced intervals, since the low frequency intervals (which would become very narrow bandwidth) might not contain any harmonic peaks.

5.2.3 Methods for Computing DCTCs Which Encode the Envelope Spectrum

The three methods described above for peak picking could each be combined with the peak DCTCs algorithm or the normal DCTC method, for a total of 6 methods for computing DCTCs which encode the envelope spectrum. In order to use the normal DCTC calculations, spectral points between peaks were first filled in by linearly interpolating between the selected peaks. This interpolated spectrum was then converted to DCTCs using both the peaks and the interpolated points. In this section we summarize the six methods. Note that the section headings are the labels used in figures and discussions for referring to these methods.

(1) Harmonic peaks

Harmonic peaks used all the harmonic peaks of the FFT spectrum and the DCTC peak algorithm to compute the DCTCs. The difficulty of this method was that the algorithm was unstable unless the number of peaks was significantly higher than the number of DCTCs to be computed. This restriction meant that for the case of the females and children, the order of the DCTC model was restricted to about 15.

(2) Harmonic peaks + linear interpolation

This method used harmonic peaks and linear interpolation to compute spectral points between harmonic peaks. Then the normal DCTC computations was performed on this envelope spectrum. However, the second method was preferred since higher order DCTC models could be reliably computed. In both cases the fundamental frequency F0 had to first be determined. Figure 5.4 shows an FFT spectrum and a spectrum recomputed from 16 DCTCs using this method.

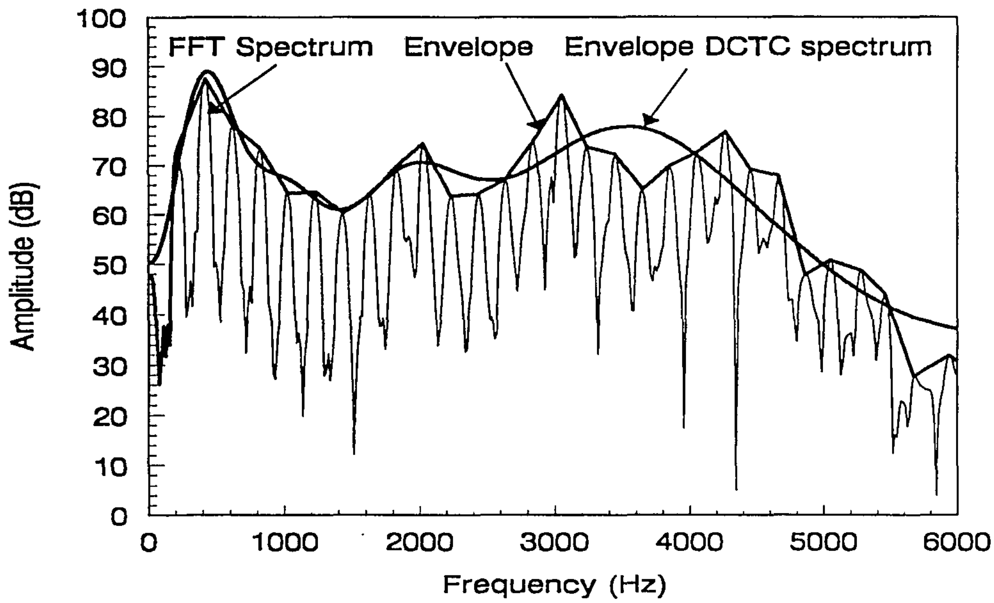

Figure 5.4 Illustration of DCTC spectrum computed using the harmonic peaks + linear interpolation method.

(3) Largest peaks in Bark-spaced

This method used only N peaks selected from the Bark-spaced intervals and the DCTC peak algorithm for computations. In the experiments, N was either 16, 18, or 20.

(4) Largest peaks in Bark-spaced + linear interpolation

In this method N peaks were selected, one each from N Bark spaced intervals. Linear interpolation filled in the spectrum between two peaks and the normal DCTC algorithm was then used to compute the DCTCs. Experiments were conducted with N values of 4, 6, 10, 12, 16, 18, 20, 25, 32, and 40. Figure 5.5 depicts spectral plots for this method for N = 16 (# of DCTCs = 16.)

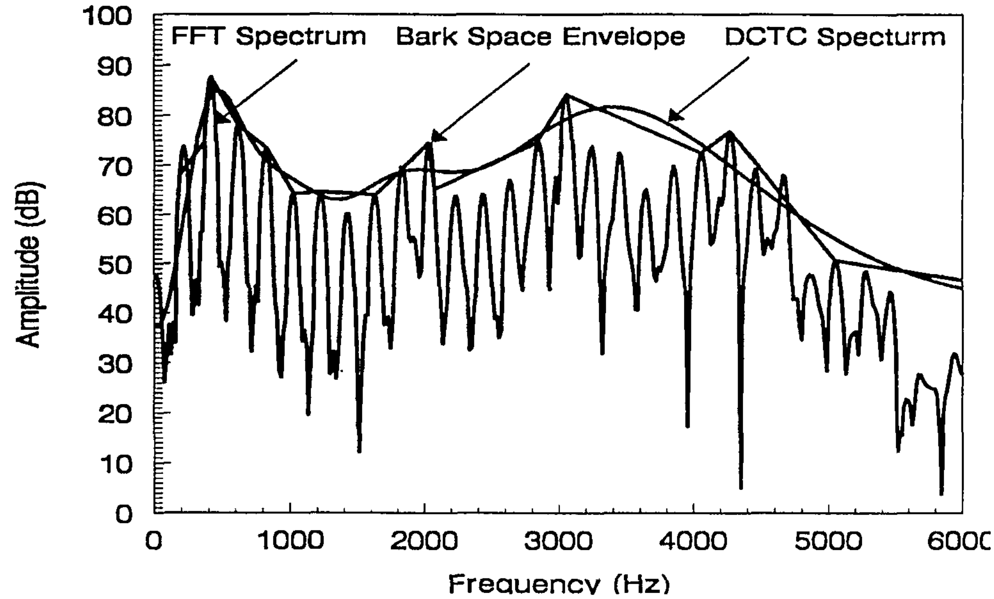

Figure 5.5 Illustration of spectrum computed from 15 envelope DCTCs computed using the largest peaks in Bark-spaced + linear interpolation method (method 4 in the text).

(5) Uniform Bark-spaced peaks

This method used the second Bark-spaced peak method described above. The DCTC peak algorithm was used to compute 16, 20, or 40 DCTCs.

(6) Uniform Bark-spaced peaks + linear interpolation

This method used the second Bark-spaced method for selecting peaks (i.e., peaks equally-spaced on a Bark scale, but chosen from an envelope spectrum with linear interpolation between harmonic peaks). The N peaks chosen as described were then again linearly interpolated and the normal DCTC algorithm was used to compute the DCTCs. Experiments were conducted for values of N equal to 6, 8, 10, 12, 16, 18, 20, 25, and 40. Figure 5.6 depicts this method using 16 Bark-spaced peaks and 16 DCTCs.

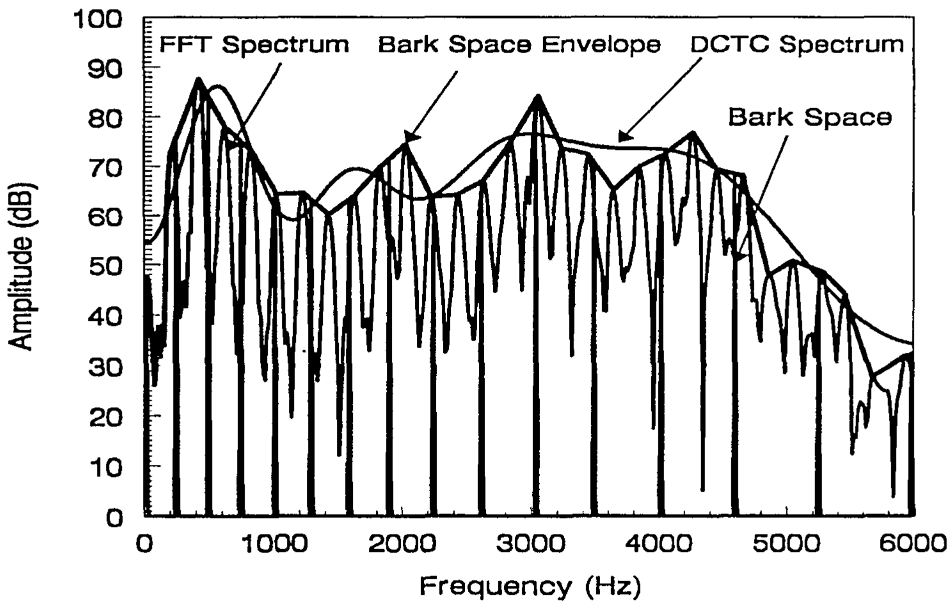

Figure 5.6 Illustration of spectrum computed from 16 envelope DCTCs computed using the uniform Bark-spaced peaks + linear interpolation method (method 6 in the text).

Figure 5.7 depicts the experimental results for these six envelope DCTC computation methods. The number of DCTCs was varied from 2 to 15. All Bark-spaced results were obtained using 20 Bark-spaced peaks. Note that the harmonic peaks + linear interpolation is the best method, and the uniform Bark space peaks + linear interpolation is second best.

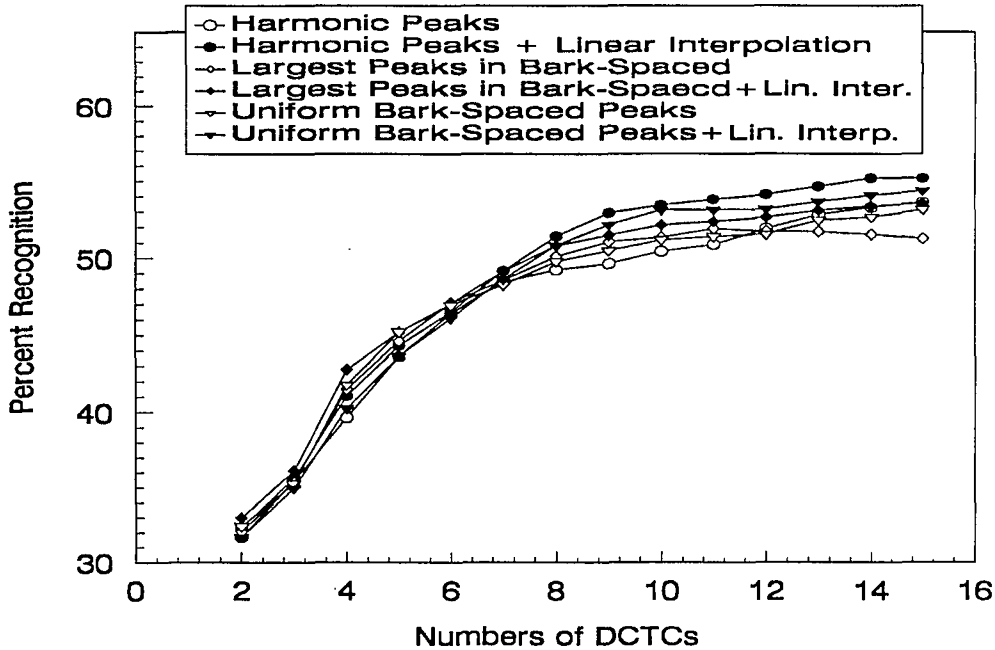

Figure 5.7 Automatic vowel classification results (13 vowels) for six envelope DCTC computation methods, as a function of the number of DCTCs used.

5.3 Primary Experiments

Having determined the values of the variables involved in the computations of the DCTC coefficients, the classification experiments were carried out. These experiments were designed to evaluate and compare normal DCTCs and envelope DCTC coefficients, as computed with the methods outlined above, via automatic classification experiments for vowels.

The data base used was the TIMIT acoustic-phonetic data base. It contains data from 630 speakers and 8 dialect regions. Each speaker read 10 sentences, for a total of 6300 sentences. For training we used 326 males and 136 females speakers, the speakers specified as training speakers on the distribution media. We used the 112 male and 56 female speakers, specified on the distribution media as test speakers, for test data. For each speaker all 10 sentences were used. Thus there were a total of 4620 sentences used for training data and 1680 sentences as testing data. The thirteen monopthongal vowels /iy, ih, ey, eh, ae, aa, ow, ah, ao, ux, uh, ax, er/ were extracted from these sentences for experimentation. The diphonthongal vowels (/ay, aw, oy/) were not used since feature trajectories should be used to classify diphthongs, and since the majority of the experiments were done using only one frame of data. Table 5.1 lists the number of vowel tokens of each type in the training set and in the test set.

Table 5.1 Number of vowel tokens in training and test sets.

	/iy/	/ih/	/ey/	/eh/	/ae/	/aa/	/ow/	/ah/	/ao	/ux/	/uh/	/ax/	/er/
Train	6668	4842	2158	3668	3733	2856	1936	2124	2795	1745	476	3285	1897
Test	2569	1604	752	1328	1278	1039	683	808	1074	525	199	1194	696

Using the computational methods for DCTCs described in the last section, several different classification experiments were conducted. For each of seven cases (six envelope DCTC methods plus normal DCTCs), fifteen DCTC coefficients were computed from one 30 ms segment selected at the labeled center of the vowel. The maximum-likelihood classifier, described previously, was used to classify vowel data with the number of DCTCs varied from 2 to 15. A portion of these experimental results are shown in figure 5.8 which shows classification rates for the three best envelope DCTC methods and the normal DCTC method. Note that the DCTCs computed from either the harmonic peaks + linear interpolation method, or with 40 uniform Bark-spaced + linear interpolation method, give results which are almost identical to the results obtained with normal DCTCs, if a large number of DCTCs are used (12 to 15). The DCTC coefficients which encode the envelope are even slightly better than the normal DCTCs if 9 to 11 DCTCs are used. Additional results, for the other DCTC methods are tabulated in Appendix B.

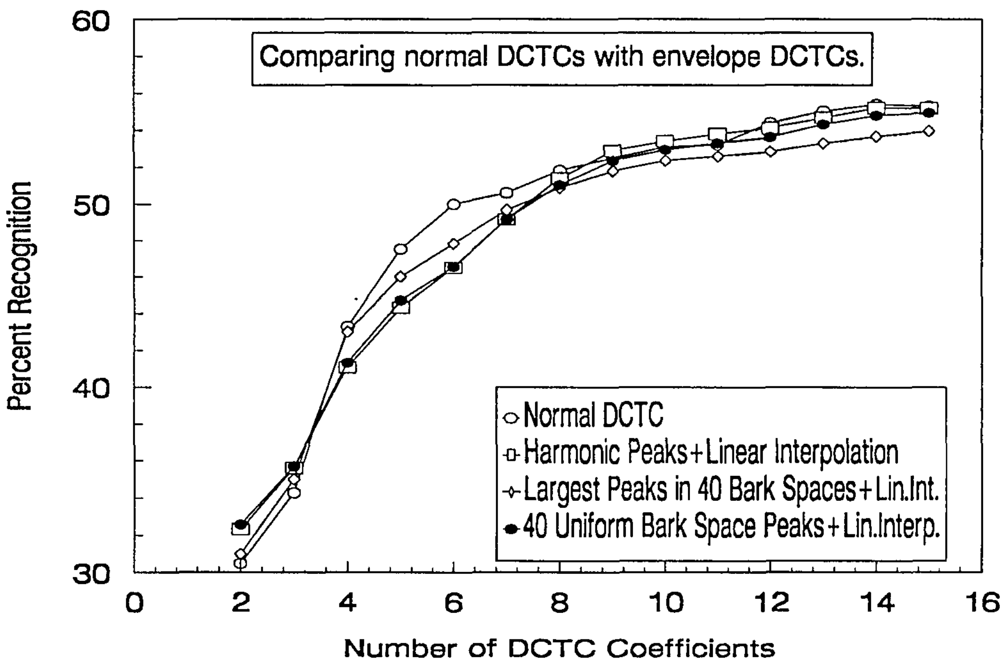

Figure 5.8 Automatic vowel classification results obtained with normal DCTCs and three types of envelope DCTCs.

Another issue, with regard to the Bark-spaced methods, was to examine the effect of the number of Bark-spaced peaks used in the calculations. This effect is illustrated in figure 5.9 for the largest peak in Bark-spaced + linear interpolation method (method 4), as the number of Bark-spaced peaks is varied from 4 to 40.

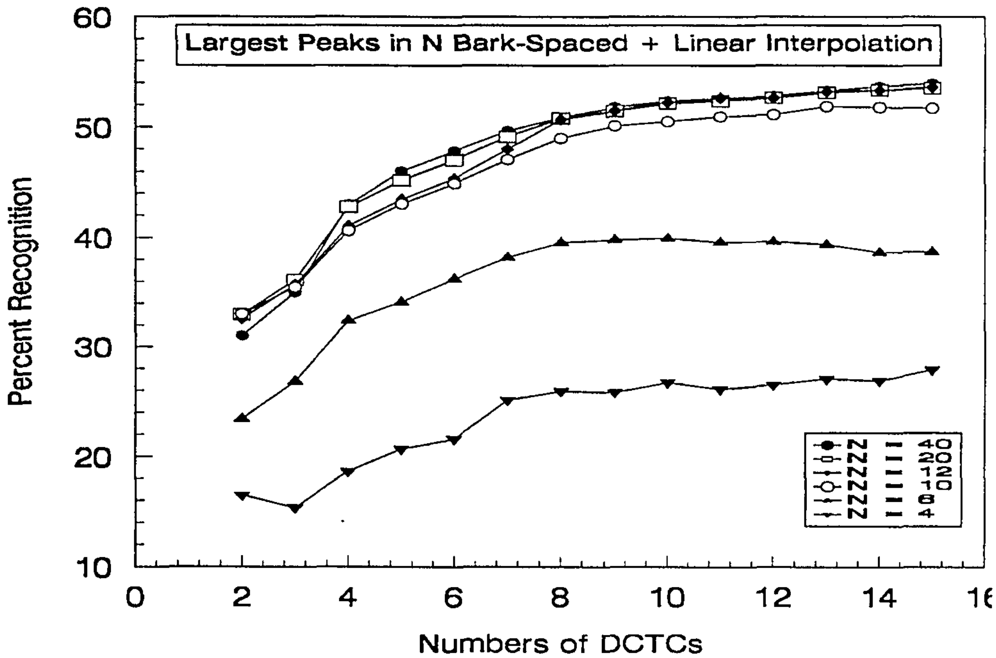

Figure 5.9 Illustration of the effect of varying the number of Bark-spaced peaks used for the largest peaks in Bark-spaced + linear interpolation method (method 4 in text).

Figure 5.10 illustrates the effect using the uniform Bark-spaced peaks + linear interpolation method (method 6 above) as the number of peaks is varied from 10 to 40. Note that with this method is unstable for N less than 10, presumably due to numerical instabilities with a matrix inverse in the Maximum likelihood classifier.

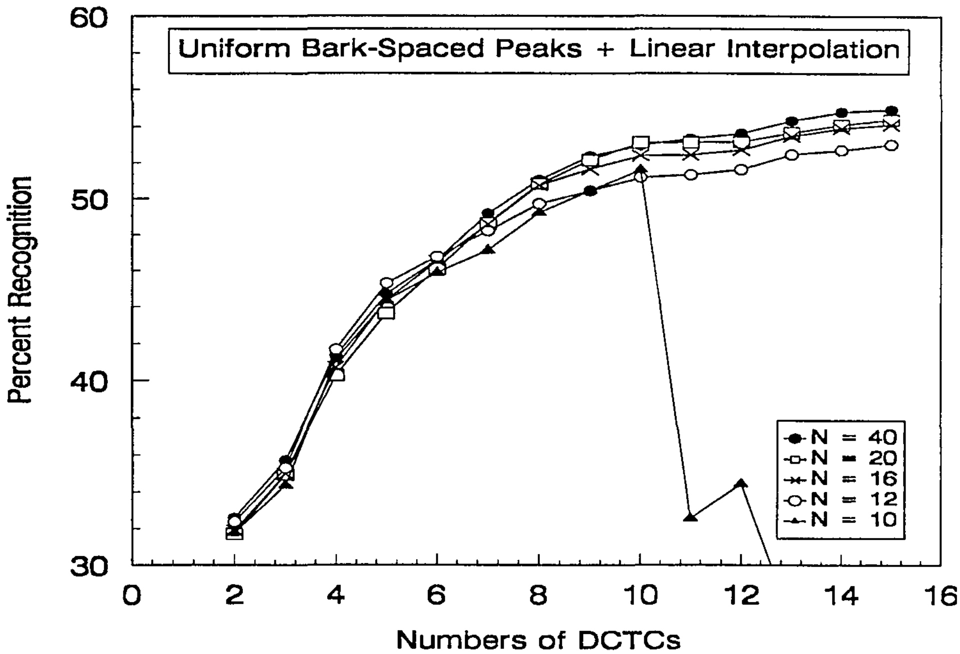

Figure 5.10 Illustration of the effect of varying the number of Bark-spaced peaks used for the uniform Bark-spaced peaks + linear interpolation method (method 6 in text).

In general these results show that classification accuracy is high if 10 or more Bark-spaced peaks are used. For the data in figure 5.9, the results are almost identical for 12 or more peaks. For the data in figure 5.10, there is a slight improvement as more peaks are added. These results show that relatively few peaks are required for the uniform Bark-spaced peaks + linear interpolation method.

Of the six methods investigated for computing the DCTCs which reflect the spectral envelope, the method based on harmonic peaks + linear interpolation gave the best results, followed by uniform Bark-spaced peaks + linear interpolation. Therefore, for additional experiments, only the harmonic peaks + linear interpolation method was used because it encodes the envelope shape of the FFT spectrum, is stable if a large number of DCTCs are used, and uses simple "normal" DCTC computations. We also note that the "best" envelope DCTCs gave equivalent performance, rather than improved performance, relative to normal DCTCs, for these classification experiments.

Despite the experimental data presented above, we thought it was still possible that the envelope DCTCs would be superior to normal DCTCs if DCTC trajectories were used for classification, as computed over several frames of speech, rather than only a single frame. Zahorian and Jagharghi (1991, 1993) have previously shown that feature trajectories can be used to improve vowel classification results. Therefore, using the method developed by Zahorian and Jagharghi, we conducted one experiment to check results based on several frames of data. For this experiment fifteen 30 ms frames were used with a 10 ms frame space. The 15 DCTCs per frame (a total of 225 features) were converted to 45 features using a 3-term cosine expansion over time for each DCTC. This method was used for both normal DCTC and envelope DCTCs. Figure 5.13 depicts the results, based on the two DCTC sets. Once again, however, the results for the normal DCTC and envelope DCTCs are essentially identical. There was no apparent improvement with features which seemed, as judged by the experimental results of the previous two chapters, to be better predictors of vowel perception.

5.4 Experiments Based on Noisy Speech

In many real world speech applications, the speech signal is corrupted by noise. Frequently, however, unlike the human perceptual system, automatic speech recognition is not particularly robust with respect to this noise. The level of the noise can be quantified by the signal to noise ratio (SNR), expressed in dB. Generally speaking when the SNR is greater then 30 dB, the signal is said to be a clean signal. If the SNR is less than -20 dB, the signal is destroyed. Intermediate values of SNR corrupt the speech signal to varying degrees. Figure 5.11 depicts a clean speech signal and noisy speech signal with a SNR = 5 dB.

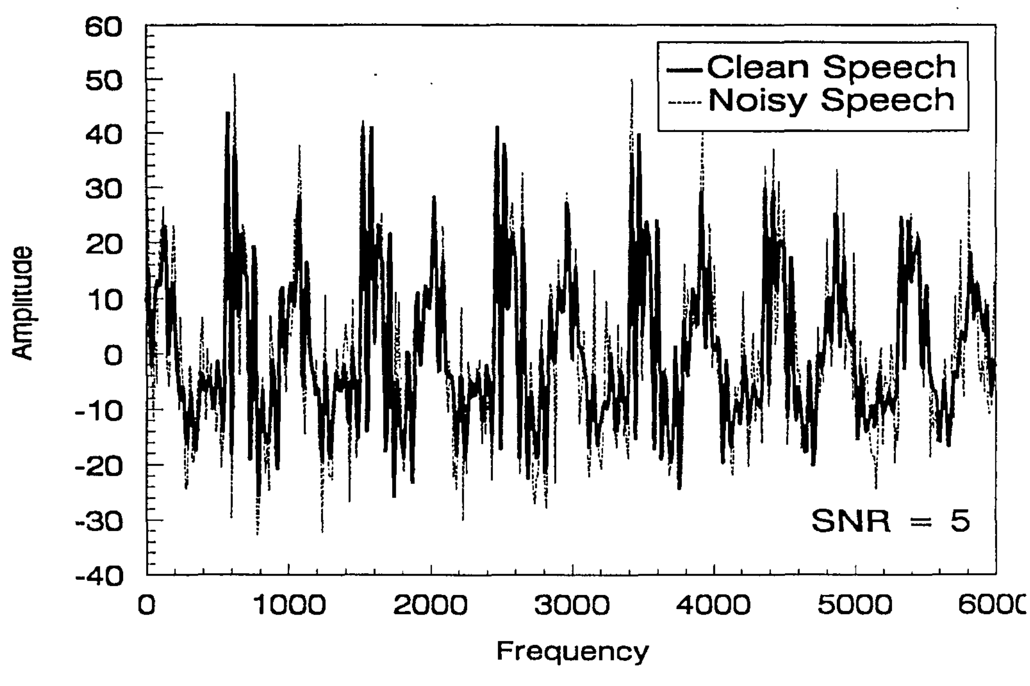

Figure 5.11 Illustration of clean speech and noisy speech (SNR = 5 dB).

We defined SNR = 20 $\log(\sigma_x^2)/(\sigma_n^2)$ for noise calculations. The noisy signal was generated by adding five independent uniform (-0.5,0.5) random variables. Therefore,σ_n^2 = 5/12 and σ_n^2 is also equals $\sigma_x^2/10^{(SNR/10)}$ from above definition. The gain of the noise was adjusted using $G = (12\sigma_n^2/5)^{1/2}$. The σ_x^2 was computed from each frame of the speech signal. The noisy speech signal can be expressed by the following equation,

$X_i = X_i + G\Sigma N_k$, where X_i is a speech sample data and N_k is a uniform random variable, $1 \leq k \leq 5$.

In this experiment we compared normal DCTCs and envelope DCTCs for vowel classification at various signal-to-noise ratios. We hypothesized that the envelope DCTCs would be better than normal DCTCs, because of the following reasoning. The normal DCT coefficients are computed using all FFT spectral components. When noise is added to the speech, the small amplitudes of FFT spectral components are more affected by the noise than are the large amplitude components (on a log amplitude scale). Therefore, the normal DCTC components will be affected to a large degree. However, since the envelope DCTC coefficients are computed using the largest amplitude peaks of the FFT spectrum, they should be much less affected by the noise. Therefore classification based on envelope DCT coefficients should remain high. Figure 5.12 depicts experimental results. For most cases the envelope DCTCs do result in higher recognition rates.

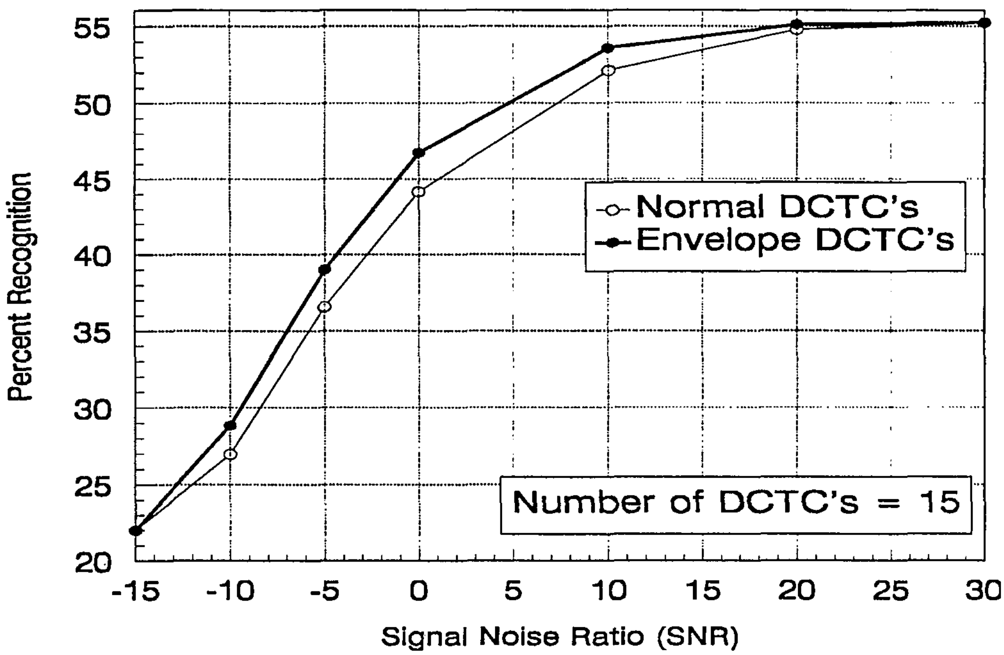

Figure 5.12 Vowel classification results for normal DCTCs and envelope DCTCs, using one frame of speech, at various signal-to-noise ratios.

The figure shows that when SNR is 30 dB, both sets of DCTCs result in the same recognition rate, which is the same as that obtained without added noise. When SNR is -15 dB, the speech information is apparently destroyed and neither set of DCTCs performs well. At intermediate SNR values, the speech is noisy, but still intelligible. At these intermediate noise levels, the envelope DCTCs outperform the normal DCTCs.

As the final vowel classification test in this series, we did a test using feature trajectories computed from noisy speech. We used 15 frames, with a 10 ms frame spacing, spanning an interval of 150 ms centered at the labeled center of each vowel token. The test was done for both clean speech and noisy speech with a SNR = 0 dB. Figure 5.13 depicts the results, which shows that envelope DCTCs are again superior to normal DCTCs in the presence of noise. Both sets of DCTCs have higher recognition rates in both clean and noisy speech, as compared with classification based on a single frame. It points out the usefulness of feature trajectories for vowel classification.

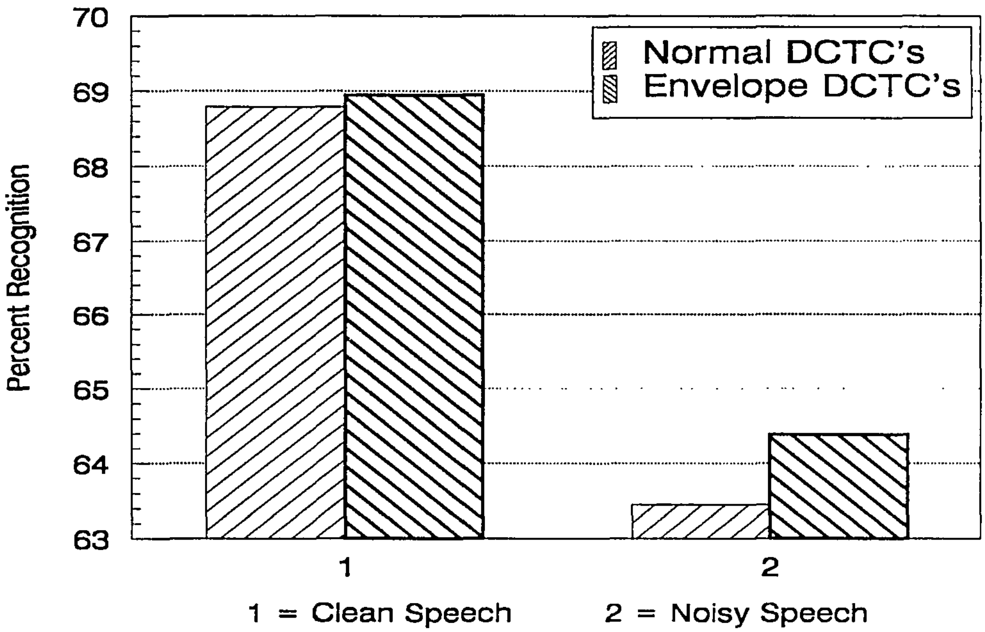

Figure 5.13 Vowel classification results obtained with DCTC trajectories for both normal DCTCs and envelope DCTCs for clean speech and noisy speech (SNR = 0 dB) with multi-frame.

The key result of this chapter is that envelope DCTCs are superior to normal DCTCs for automatic vowel classification, if classification is based on noisy speech, and nearly identical for the case of clean speech. The similar performance of the two features sets for clean speech is undoubtedly because for clean naturally-produced speech (unlike the reduced harmonic speech used for synthesis), the normal DCTCs and envelope DCTCs reflect nearly identical spectral properties. Figure 5.14 shows both DCTC spectrums for a natural speaker. However, the improvement obtained with using envelope DCTCs for the case of noisy speech is potentially important in improving the robustness of automatic speech recognition.

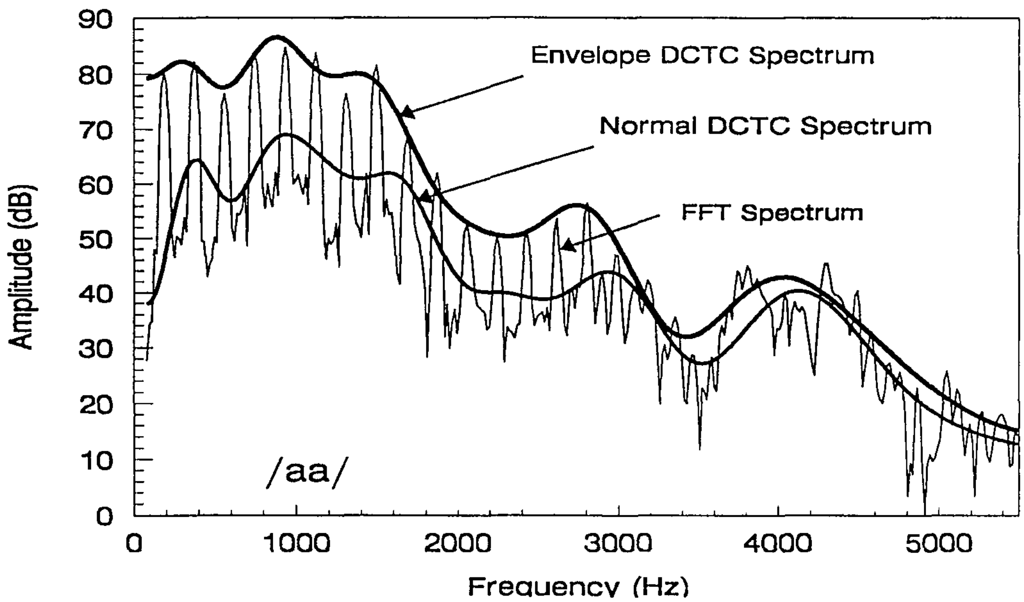

Figure 5.14 Illustration of normal DCTC spectrum and envelope DCTC spectrum for a natural speaker.

CHAPTER SIX

Conclusion

In this study several issues related to acoustic correlates of vowel perception were investigated using speech synthesis of multi-tone stimuli, using various criteria for selecting the amplitudes and frequencies of the tones, and the perception of these stimuli, as a test bed. We first compared two already-developed feature sets as acoustic correlates. In particular, we compared vowel perception of multi-tone stimuli synthesized to either primarily preserve formant frequencies or to preserve DCTC spectral shape features. The results of the experiments imply the following points.

1. Vowel stimuli which preserve fromant frequencies, but which distort spectral shape, are perceptually impoverished. In contrast, vowel stimuli which preserve spectral shape, but which only approximately preserve formants, are identified with greater accuracy.

2. The largest peaks of the spectrum play an important role in synthesized vowels. It led us to examine how many of the largest peaks are required to synthesize vowels such that multi-tone stimuli are well identified. From our experiments, depending on the vowel, between 5 and 10 sinusoids are required such

that a synthesized token is identified with nearly the same accuracy as the original vowel. However, careful listening does show that the vowels synthesized with reduced harmonics are still perceived as sounding different from the original token.

Our first series of tests demonstrated that formants are insufficient cues for vowel perception, and that a more complete set of cues are required to reliably predict perception. Although the spectral features originally investigated do provide a more complete spectral description, we also determined that it is possible to synthesize vowel tokens which preserve these spectral shape features but which do not preserve vowel perception. This conclusion led us to the second phase of the work, to reformulate a definition of spectral shape features which would be more consistent with perceptual results.

Therefore the definition of the DCT coefficients was modified so that the DCTCs represented an encoding of the envelope spectrum. That is, a spectrum recomputed from these DCTCs is a smoothed version of the envelope of the original FFT spectrum. This redefinition of the DCTC spectral shape factors was motivated by the first experiment which indicated that multi-tone stimuli should preserve the largest peaks in the spectral envelope to preserve vowel intelligibility. We investigated these new DCTCs both with perceptual tests, again using the sinusoidal synthesizer, and with automatic classification experiments.

The next conclusion of our work, as obtained from the second experiment, is the following:

3. Synthesis of multi-tone stimuli which preserves the peak DCTCs, thus preserving the envelope spectrum, results in much higher vowel intelligibility than is obtained from tokens synthesized to preserve the original DCTCs. Thus the peak-derived DCTCs are much more viable as acoustic correlates of vowel perception than are the normal DCTCs.

In the last experiment, we tested the effectiveness of envelope DCTCs for automatic vowel classification, as opposed to classification based on normal DCTCs. The results of this experiment are summarized as point 4.

4. The envelope DCT coefficients and normal DCTC coefficients result in almost identical vowel classification rates if the speech is noise free. However, the envelope DCTC coefficients result in higher classification rates than do normal DCTCs if the speech signal is corrupted by noise.

In terms of the overall objective, to formulate a set of acoustic correlates for vowel perception, our study was partially successful. We did observe that formants alone are insufficient correlates. We also determined that perception gradually improves as more and more spectral detail is added. This result implies acoustic correlates are required which encode the entire spectrum. The best correlates found were the envelope DCTCs. However, even these correlates are not completely consistent with the major result found from the sinusoidal synthesis and perceptual experiments. Namely, the best approach to maximize vowel intelligibility with a fixed number of sinusoids is to use sinusoids corresponding to the largest spectral peaks. This strategy is not equivalent to choosing sinusoids which best preserve the envelope DCTCs.

An issue for future study is to further improve the formulation of the global spectral shape acoustic correlates. These correlates should be good indicators of perception and also be useful for automatic speech recognition in both clean and noisy speech. Such a new feature set for representing the speech signal will have important consequences for speech signal processing. Another point which might be investigated in more detail is a more sophisticated scheme for selecting spectral peaks from frame to frame in multi-frame speech in the presence of noise. In particular, since the noise is uncorrelated over time (assuming white noise), whereas the speech signal harmonics are continuous in time, phase considerations in the spectrum might be used to better separate speech from noise.

BIBLIOGRAPHY

Duda, R. O. and Hart, P. E. (1973). Pattern Analysis and Scene Classification, Wiley & Sons, New York.

Effer, E. A. (1985). "An investigation to improve linear predictive vocoder pulse excitation models," Master's thesis, Old Dominion University.

Fant, G. (1960). Acoustic Theory of Speech Production, (Mouton and Co.'s. Gravenhage, The Netherlands).

Jagharghi, A. J. (1990) "A comparative study of spectral peaks versus global spectral shape as invariant acoustic cues for vowels," Ph.D. Dissertation, Old Dominion University.

Jagharghi, A. J. and Zahorian, S. A. (1990). "Vowel perception:Spectral shape versus formants," J.Acoust.Soc. Amer. 87, S159.

James Mike (1985). Classification Algorithms, Collins.

Kakusho, O., Hirato, H. and Kato K. (1971). "Some Experiments of Vowel Perception by Harmonic Synthesizer," Acustica. Vol. 24, 1971.

Kates, J. M. (1992). "Speech enhancement based on a sinusoidal model," City University of New York, 1992.

Klatt, D. H. (1980)."Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Amer. 67, 971-995.

Markel, J. D. (1972). "The Sift algorithm for fundamental frequency estimation," IEEE Trans. Audio & Electroacoustics 20, 367-377.

McAulay, R. J. and Quatieri, T. F. (1986). "Speech analysis/synthesis based on a sinusoidal representation," IEEE Trans. Acoust. Speech and Sig. Proc., Vol. ASSP-34(4), 744-754.

Nossair, Z. B. and Zahorian, S. A. (1991). "Dynamic spectral shape feature as acoustic correlates for initial stop consonants," J.Acoust.Soc. Amer. 89, 2978-2991.

Oppenheim, A. V. and Schafer, R. W. (1989). Discrete-Time Signal Processing, (Prentice-Hall, New Jersey, 1989).

Pols, L. C. W. (1977)."Spectral analysis and identification of Dutch vowels in monosyllabic words," Institute for perception TNO,(Solsterberg, The Netherlands.)

Remez, R. E. and Pison, D. B. (1981). "Speech Perception Without Traditional Speech Cues," Science Vol. 212, 22.

Sydal, AnnK. and Gopal, H. S. (1986)."A Perceptual Model of Vowel Recognition Based on the Auditory Representation of American English Vowels," J.Acoust.Soc.Amer. Vol. 79(4)

Zahorian, S. A. and Gordy, P. E. (1983). "Finite impulse response (FIR) filters for speech analysis and synthesis," ICASSP-83, 808-811.

Zahorian, S. A. and Jagharghi, A. J.(1991). "Speaker normalization of static and dynamic vowel spectral feature," J.Acoust.Soc. Amer. 90, 67-75.

Zahorian, S. A. and Jagharghi, A. J.(1993). "Spectral shape features versus formants as acoustic correlates for vowel," J.Acoust.Soc. Amer. 1993.

Zahorian, S. A. and Rudasi, L. (1993). "Frequency warping with modified cosine transform basis functions."

Zwicker, E. (1961). "Subdivision of audible frequency range into critical band (Frequenzgruppen)," J. Acoust. Soc. Amer. 33, 248.

Appendix A.

Confusion Matrix of Listening Experiments for Chapter 4.

Table A.1 The confusion matrix in terms of percentage for Original Speech.

	/aa/	/iy/	/uw/	/ae/	/er/	/ih/	/eh/	/ao/	/ah/	/uh/
/aa/	100									
/iy/		89				11				
/uw/			100							
/ae/				56			33		11	
/er/					100					
/ih/						34	66			
/eh/				33			67			
/ao/	55							45		
/ah/	11						11		78	
/uh/								11	44	45

Table A.2 The confusion matrix in terms of percentage for One Period Repetition.

	/aa/	/iy/	/uw/	/ae/	/er/	/ih/	/eh/	/ao/	/ah/	/uh/
/aa/	100									
/iy/		100								
/uw/			89							11
/ae/				78			22			
/er/				11	78		11			
/ih/		22				56	22			
/eh/				66		11	23			
/ao/	44							45	11	
/ah/	22							11	67	
/uh/									44	56

Table A.3 The confusion matrix in terms of percentage for All Original Peaks.

	/aa/	/iy/	/uw/	/ae/	/er/	/ih/	/eh/	/ao/	/ah/	/uh/
/aa/	89			11						
/iy/		100								
/uw/			89							11
/ae/	11			89						
/er/					100					
/ih/						23	77			
/eh/	11			66			23			
/ao/	44							56		
/ah/	33							22	45	
/uh/								33	33	34

Table A.4 The confusion matrix in terms of percentage for 5 Original Peaks.

	/aa/	/iy/	/uw/	/ae/	/er/	/ih/	/eh/	/ao/	/ah/	/uh/
/aa/	56			22	11	11				
/iy/		100								
/uw/			78							22
/ae/			11	23				44		22
/er/					78		11		11	
/ih/			22		11	45	22			
/eh/				11			0	77		12
/ao/	22		11	11				45	11	
/ah/		11					11		45	33
/uh/									11	89

Table A.5 The confusion matrix in terms of percentage for Formats.

	/aa/	/iy/	/uw/	/ae/	/er/	/ih/	/eh/	/ao/	/ah/	/uh/
/aa/	11								22	
/iy/		78								
/uw/			88							11
/ae/		11		0	11	33	22		22	
/er/					55	22	11			11
/ih/		33				33	22			11
/eh/		22				44	33			
/ao/					11	11	11	0	44	22
/ah/				11	44	11		11	11	11
/uh/	11		33		11				11	33

Table A.6 The confusion matrix in terms of percentage for Formants Plus Side Peaks.

	/aa/	/iy/	/uw/	/ae/	/er/	/ih/	/eh/	/ao/	/ah/	/uh/
/aa/	33			22		11		11	22	
/iy/		100								
/uw/			88							
/ae/		11		56			33			
/er/					78	11	11			
/ih/		11				56	33			
/eh/	11	11				11	56			
/ao/	22							56	11	11
/ah/									78	22
/uh/						11	11		33	45

Table A.7 The confusion matrix in terms of percentage for 4 Largest Peaks Plus One Side Peaks.

	/aa/	/iy/	/uw/	/ae/	/er/	/ih/	/eh/	/ao/	/ah/	/uh/
/aa/	100									
/iy/		100								
/uw/			89							11
/ae/	11			45			33		11	
/er/					89				11	
/ih/		11				56	33			
/eh/	11			55			23	11		
/ao/	44			11				45		
/ah/	11			11	33			11	34	
/uh/			11		33			11	11	34

Table A.8 The confusion matrix in terms of percentage for 16 Bark Equeal Space.

	/aa/	/iy/	/uw/	/ae/	/er/	/ih/	/eh/	/ao/	/ah/	/uh/
/aa/	11			22				56	11	
/iy/		89				11				
/uw/			78		11					11
/ae/		11		56.			22	11		
/er/			22		78					
/ih/			11	11		56	22			
/eh/		11	22		11	44				11
/ao/					11			56	22	11
/ah/	33		11						34	22
/uh/			33		11				56	0

Table A.9 The confusion matrix in terms of percentage for 8 Peaks Out of 4 Largest Peaks.

	/aa/	/iy/	/uw/	/ae/	/er/	/ih/	/eh/	/ao/	/ah/	/uh/
/aa/	11		11		45			11	11	11
/iy/		100								
/uw/	22				11			33	11	22
/ae/	11	11	11	11		33	22			
/er/			22		56	11	11			
/ih/		44	11			33	11			
/eh/			11			33	44		11	
/ao/			11						44	44
/ah/	22			11	33				33	
/uh/	11		22	11	22		11	11	11	

Table A.10 The confusion matrix in terms of percentage for 16 FFT Peaks.

	/aa/	/iy/	/uw/	/ae/	/er/	/ih/	/eh/	/ao/	/ah/	/uh/
/aa/	89								11	
/iy/		100								
/uw/			56		11					33
/ae/				89			11			
/er/				11	89					
/ih/						45	55			
/eh/				22			78			
/ao/	33							56		11
/ah/	44							22	34	
/uh/			11	11	11					67

Table A.11 The confusion matrix in terms of percentage for 16 DCTC Peaks.

	/aa/	/iy/	/uw/	/ae/	/er/	/ih/	/eh/	/ao/	/ah/	/uh/
/aa/	89			11						
/iy/		100								
/uw/			89							11
/ae/				89		11				
/er/			11		67		22			
/ih/						45	55			
/eh/				33			67			
/ao/	33							56		11
/ah/	22							44	33	
/uh/				11		11				78

Appendix B.

The Experiments Results for Classification

Table B.1 Results of classification for one frame.

Normal DCTCs			DCTC peak (Harmonic)			Envelope DCTC (Harmonic+Lin.Inter.)		
#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)
02	30.668	30.497	02	31.414	32.097	02	31.517	32.330
03	34.531	34.293	03	34.476	35.799	03	34.662	35.617
04	42.816	43.276	04	39.804	40.672	04	40.762	41.072
05	46.959	47.531	05	43.533	44.345	05	43.525	44.316
06	49.618	49.982	06	46.524	46.949	06	46.056	46.505
07	50.519	50.593	07	48.544	48.978	07	48.696	49.153
08	51.548	51.822	08	49.306	49.385	08	51.173	51.386
09	52.386	52.484	09	50.479	50.287	09	52.952	52.891
10	52.752	53.085	10	51.517	50.826	10	54.002	53.400
11	53.191	53.209	11	52.530	51.618	11	54.335	53.793
12	54.359	54.384	12	53.397	52.797	12	54.984	54.113
13	54.879	55.027	13	54.230	53.531	13	55.409	54.659
14	55.118	55.392	14	54.623	53.771	14	55.587	55.197
15	55.173	55.297	15	54.937	53.997	15	55.671	55.175

Table B.2 Largest peaks in Bark-spaced selected.

N Bark Space-spaced + Linear Interpolations

N = 40			N = 32			N = 25		
#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)
02	30.799	31.006	02	31.461	32.141	02	32.145	33.195
03	34.408	35.021	03	34.825	36.032	03	35.356	36.446
04	42.672	42.992	04	42.569	43.261	04	42.931	43.050
05	45.621	46.011	05	45.534	45.967	05	45.684	46.047
06	47.776	47.843	06	47.782	47.902	06	47.541	47.974
07	49.251	49.669	07	49.026	49.102	07	49.054	49.444
08	50.767	50.862	08	50.592	50.215	08	50.663	50.636
09	52.095	51.800	09	51.915	51.371	09	52.163	51.153
10	52.968	52.346	10	52.706	52.142	10	52.944	52.186
11	53.243	52.578	11	52.902	52.069	11	52.829	52.375
12	53.858	52.833	12	53.557	52.593	12	53.628	52.513
13	54.285	53.284	13	54.041	53.058	13	54.209	53.131
14	54.408	53.655	14	54.382	53.277	14	54.335	53.233
15	54.607	53.946	15	54.856	53.517	15	54.856	53.735

(continue Table B.2)

N = 20			N = 18			N = 16		
#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)
02	32.067	32.984	02	31.883	32.788	02	31.718	32.548
03	35.503	36.134	03	35.100	36.097	03	35.034	36.119
04	42.423	42.759	04	41.930	42.119	04	41.393	41.930
05	45.233	45.189	05	44.783	45.080	05	44.309	44.520
06	46.988	47.036	06	46.430	46.913	06	45.927	46.447
07	48.735	49.131	07	48.308	48.956	07	48.298	48.869
08	50.561	50.789	08	50.241	50.658	08	50.453	50.702
09	52.116	51.451	09	51.726	51.327	09	51.776	51.684
10	52.672	52.113	10	52.342	52.077	10	52.548	52.215
11	52.779	52.353	11	52.800	52.644	11	53.025	52.687
12	53.531	52.629	12	53.201	52.629	12	53.345	52.695
13	54.162	53.095	13	53.704	53.248	13	53.735	53.248
14	54.233	53.349	14	53.950	53.320	14	54.023	53.582
15	54.515	53.524	15	54.497	53.488	15	54.361	53.771

(continue Table B.2)

N = 12			N = 10			N = 6			N = 4		
#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)
02	32.006	32.613	02	31.920	33.057	02	22.931	23.405	2	16.176	16.496
03	34.987	35.704	03	34.594	35.508	03	26.273	26.795	3	14.827	15.354
04	40.707	41.036	04	40.244	40.636	04	31.970	32.417	4	18.638	18.656
05	43.295	43.465	05	42.548	43.036	05	33.630	34.141	5	20.278	20.678
06	45.495	45.363	06	44.450	44.898	06	35.982	36.250	6	21.160	21.616
07	48.057	48.018	07	46.852	47.080	07	38.287	38.199	7	24.678	25.085
08	50.330	50.702	08	48.520	49.014	08	39.209	39.567	8	25.673	25.929
09	51.472	51.473	09	50.267	50.105	09	39.120	39.836	9	26.066	25.798
10	52.087	52.229	10	50.642	50.491	10	39.261	39.945	10	26.933	26.693
11	52.719	52.629	11	50.953	50.956	11	38.832	39.603	11	26.383	26.089
12	52.724	52.629	12	51.158	51.131	12	39.002	39.647	12	27.087	26.489
13	53.211	53.218	13	51.566	51.858	13	39.091	39.399	13	27.530	27.042
14	53.578	53.349	14	51.818	51.793	14	38.594	38.679	14	27.255	26.860
15	53.727	53.618	15	51.928	51.713	15	38.389	38.759	15	27.960	27.900

Table B.3 Uniform Bark-Spaced peaks selected.

Uniform Bark-spaced amplitudes base on envelope								
N = 40			N = 20			N = 16		
#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)
02	31.716	32.911	02	31.108	32.431	02	30.938	32.097
03	34.804	35.966	03	34.295	35.275	03	34.107	34.948
04	41.351	41.683	04	41.310	41.778	04	39.683	40.337
05	44.041	44.687	05	44.096	45.218	05	43.038	43.698
06	46.543	46.694	06	46.228	46.934	06	45.440	45.931
07	48.541	48.964	07	48.153	48.236	07	46.899	47.545
08	50.616	50.680	08	49.895	49.756	08	48.080	48.273
09	52.247	51.778	09	50.951	50.455	09	49.136	49.313
10	53.038	52.513	10	51.632	51.160	10	50.107	50.062
11	53.452	52.811	11	51.763	51.371	11	51.032	50.549
12	54.102	53.335	12	52.596	51.516	12	51.598	51.400
13	54.838	54.062	13	53.167	52.440	13	51.902	51.633
14	55.094	54.564	14	53.318	52.629	14	52.551	51.975
15	55.610	54.928	15	53.937	53.138	15	53.098	52.600

Table B.4 Uniform Bark-Spaced peaks + Linear Interpolations.

Uniform Bark-spaced peaks + Linear Interpolations.

N = 40			N = 25		
#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)
02	31.624	32.591	02	31.375	32.162
03	34.772	35.712	03	34.437	35.137
04	41.016	41.327	04	40.257	40.628
05	43.866	44.716	05	42.933	43.770
06	46.404	46.571	06	45.592	46.120
07	48.628	49.160	07	48.465	49.153
08	50.943	51.029	08	50.896	51.655
09	52.504	52.353	09	52.535	52.542
10	53.345	52.927	10	53.625	53.218
11	53.777	53.298	11	54.060	53.633
12	54.518	53.611	12	54.508	54.258
13	55.105	54.302	13	55.045	54.419
14	55.254	54.789	14	55.314	54.957
15	55.576	54.920	15	55.461	54.731

(Continue Table B.4)

N = 20			N = 18			N= 16		
#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)
02	30.922	31.704	02	31.273	32.031	02	30.961	31.901
03	34.065	34.955	03	34.445	35.290	03	34.083	34.992
04	40.427	40.257	04	40.241	40.541	04	40.597	40.694
05	43.292	43.640	05	43.080	43.785	05	43.444	44.352
06	45.982	46.076	06	45.458	46.083	06	45.980	46.505
07	48.321	48.607	07	48.229	48.418	07	48.193	48.571
08	50.545	50.811	08	50.550	50.789	08	50.102	50.738
09	52.153	52.127	09	51.977	51.662	09	51.692	51.626
10	53.080	53.087	10	52.831	52.520	10	52.478	52.375
11	53.470	53.102	11	53.342	52.884	11	52.816	52.411
12	54.023	53.138	12	53.628	53.066	12	53.318	52.702
13	54.633	53.640	13	54.062	53.553	13	53.892	53.444
14	54.730	54.069	14	54.615	54.048	14	54.311	53.880
15	55.084	54.360	15	54.869	54.215	15	54.460	54.077

(continue Table B.4)

N = 12			N = 10		
#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)
02	31.092	32.388	02	30.285	31.770
03	34.337	35.304	03	33.667	34.366
04	41.281	41.661	04	40.524	41.108
05	44.117	45.291	05	43.457	44.345
06	46.126	46.752	06	45.542	45.872
07	48.193	48.178	07	47.177	47.109
08	49.762	49.691	08	49.597	49.189
09	50.922	50.411	09	51.160	50.425
10	51.579	51.160	10	52.402	51.582
11	51.815	51.298	11	32.750	32.562
12	52.517	51.611	12	34.544	34.446
13	53.193	52.440	13	26.797	27.049
14	53.381	52.687	14	20.230	20.336
15	53.800	52.993	15	19.618	18.940

Table B.5 Comparison normal DCTC's and envelope DCTC's in various SNR.

Add noise to signal in normal (sum 5 uniform (-.5, +.5)) 13 vowels BTWC = .45

normal DCTC			envelope DCTC		
SNR = 30					
#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)
02	30.118	30.199	02	32.669	33.399
03	33.869	33.755	03	36.194	37.195
04	43.057	43.538	04	41.592	41.588
05	47.116	47.574	05	44.112	44.381
06	49.722	49.945	06	46.530	46.796
07	50.935	51.044	07	49.034	49.211
08	52.056	52.593	08	51.003	51.211
09	52.981	53.357	09	52.724	52.542
10	53.918	54.229	10	53.502	53.451
11	54.646	54.899	11	54.170	53.982
12	55.220	55.284	12	54.717	54.251
13	55.385	55.233	13	55.136	54.528
14	55.647	55.146	14	55.280	55.197
15	55.691	55.233	15	55.414	55.240

(continue Table B.5)

SNR = 20

#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)
02	31.133	31.089	02	32.910	33.508
03	34.058	34.828	03	37.061	37.574
04	41.657	41.058	04	42.208	42.563
05	43.105	41.333	05	44.610	44.687
06	46.805	46.544	06	47.080	46.869
07	48.685	48.893	07	49.112	49.291
08	49.980	50.033	08	50.862	50.746
09	51.746	51.944	09	52.564	52.207
10	52.837	53.011	10	53.185	52.978
11	53.664	53.960	11	53.751	53.691
12	54.288	54.091	12	54.466	53.953
13	54.770	54.295	13	54.822	54.528
14	54.971	54.469	14	54.935	54.695
15	55.170	54.760	15	55.168	55.095

(continue Table B.5)

SNR = 10

#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)
02	31.839	32.786	02	31.532	32.286
03	35.947	37.829	03	36.048	36.628
04	41.057	41.367	04	41.854	42.258
05	42.933	43.013	05	44.405	44.323
06	44.635	45.016	06	46.205	46.040
07	46.244	46.679	07	47.645	47.873
08	47.840	48.542	08	49.180	49.233
09	49.066	49.424	09	50.736	50.646
10	50.342	50.682	10	51.558	51.285
11	51.084	50.891	11	52.040	51.702
12	51.587	51.407	12	52.287	51.923
13	52.001	51.720	13	52.850	52.682
14	52.488	51.793	14	53.339	53.105
15	52.734	52.113	15	53.672	53.554

(continue Table B.5)

SNR = 0

#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)
02	27.624	28.075	02	27.148	27.442
03	30.267	30.599	03	31.496	31.770
04	35.770	36.344	04	36.391	37.450
05	38.009	38.352	05	38.494	39.065
06	39.222	39.443	06	39.348	39.457
07	40.605	40.236	07	40.914	41.298
08	41.351	41.043	08	42.782	42.839
09	42.520	42.003	09	44.120	44.330
10	43.114	42.694	10	44.759	44.963
11	43.442	43.283	11	45.157	45.058
12	43.751	43.443	12	45.783	45.691
13	44.353	43.807	13	46.359	46.003
14	44.536	43.829	14	46.839	46.520
15	44.890	44.170	15	47.386	46.745

(continue Table B.5)

SNR = -5

#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)
02	24.361	24.656	02	23.960	24.387
03	25.917	25.922	03	26.820	27.231
04	29.893	30.293	04	30.739	31.391
05	30.959	31.573	05	31.998	32.279
06	32.101	32.286	06	32.829	32.991
07	33.229	33.115	07	34.217	34.002
08	33.937	33.704	08	35.498	35.115
09	34.670	34.242	09	37.067	36.504
10	35.382	35.283	10	37.645	37.254
11	35.733	35.173	11	38.038	37.552
12	36.008	35.203	12	38.408	37.799
13	36.537	35.719	13	39.062	38.105
14	36.781	36.039	14	39.547	38.752
15	37.190	36.563	15	39.997	38.999

(continue Table B.5)

SNR = -10

#DCTC	Train(%)	Test(%)	#DCTC	Train(%)	Test(%)
02	21.891	22.045	02	21.799	21.936
03	22.292	22.365	03	22.763	23.245
04	24.099	24.555	04	24.618	25.245
05	24.398	24.700	05	25.071	25.180
06	25.149	25.195	06	25.705	25.507
07	25.783	25.507	07	26.427	26.140
08	26.118	25.573	08	27.250	26.664
09	26.514	25.878	09	28.043	27.202
10	26.771	26.242	10	28.586	27.595
11	27.085	26.336	11	28.724	28.213
12	27.349	26.242	12	29.099	28.249
13	27.750	26.773	13	29.646	28.409
14	27.931	26.984	14	30.063	28.686
15	28.321	26.962		30.422	28.846

Table B.6 comparison of normal DCTC and envelope DCTC for Multi-frames in clean and noisy speech signal.

15 frame (30ms frame length and 10 ms-frame spacing)

	Normal DCTC		Envelope DCTC	
	Train	Test	Train	Test
15 DCTCs per frame	68.8	66.2	69.0	65.9
12 DCTCs per frame	67.6	65.3	67.8	64.9
15 DCTC per frame+noise(SNR=0)	63.5	59.4	64.4	60.9