COMPUTER RECOGNITION OF STOP CONSONANTS IN CONTINUOUS SPEECH USING DIFFERENT DISTINCTIVE FEATURES

Blevins Tang

A Thesis
in
The Department
of
Computer Science

Presented in Partial Fulfillment of the Requirements
for the degree of Master of Computer Science at
Concordia University
Montreal, Quebec, Canada

September 1981

© Blevins Tang, 1981

ABSTRACT

COMPUTER RECOGNITION OF STOP CONSONANTS IN CONTINUOUS SPEECH USING DIFFERENT DISTINCTIVE FEATURES

Blevins Tang

To distinguish voiced and unvoiced stop consonants, many models using different single features have been developed by a number of researchers. Of all these models, a complete computer system to recognize stops, /p,t,g,b,d,k/, in continuous speech has not yet been established. Most of the researchers rely on features measured from sonograms. In this thesis the author proposes a completely automatic speaker independent system to recognize the stops in continuous speech. This system makes use of three different distinctive features which include formant transitions, silent interval and voice onset time. Data were collected from four male and four female English speakers. Each of them uttered twenty sentences of medium size. Altogether, 320 sentences were recorded of which only 120 sentences from six speakers with highest scores were selected and tested by the system. The results confirm that no single feature can account for the distinction of voiced and unvoiced stop consonants. A comparison of the results of three different distinctive features has been made. Comparatively, the formant transitions provide the best cue among the three features.

The respective average recognition rates are 70.58 percent for unstressed data and 79.56 percent for stressed data for the six speakers.

ACKNOWLEDGEMENTS

The author would like to express his grateful appreciation to his supervisor, Dr. C. Y. Suen, Department of Computer Science of Concordia University, for his excellent guidance, valuable advice, suggestions and patience during the preparation of this thesis.

Special acknowledgement and thanks go to the speakers, the sound studio operator and J. Mulherin, especially the latter who set up the hardware system for data digitization.

Also thanks have to be given to the staff of the Computer Center of Concordia University for their cooperation, and the colleagues for their comments and suggestions.

This research was financially supported by the Department of Education of Quebec.

- * TABLE OF CONTENTS * -

LIST OF FIGURES

Figure Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

I.1 SPEECH RECOGNITION PHILOSOPHY

Automatic speech recognition (ASR) is regarded as the most important tool in the field of communication between human and machine. A considerable amount of ASR research work has been done by many researchers for more than three decades.

If we look back at the development of ASR, we could find the original book referring to speech recognition written by the pioneers Potter, Kopp and Kopp in 1947[1]. They created application of electronic instrumentation to speech processing. The early papers on ASR were written by Dreyfus-Grof[2] in 1950, Davis et. al.[3] in 1952, and Fry and Denes[4] in 1953. They claimed their work could achieve high recognition scores for a dictionary of about ten words. These words, which were 10 numerals, had to be spoken in isolation by a single speaker whose voice the machine was adjusted to. However, considerably lower scores were obtained when several speakers were used. In the nineteen sixties, extensive research efforts brought noticeable improvements to this area. ASR devices were developed which could recognize 90% of twenty to fifty isolated words spoken by several speakers. Simple continous speech such as phrases or short sentences were also recognized under laboratory conditions. Towards the end of the sixties, research on the recognition of continuous speech

In ASR, a feasible way of recognizing speech is to recognize the phonemes instead of the entire words because phonemes are limited in number. For instance, in American English, there are only 42 phonemes[5] (12 vowels, 6 diphthongs, 4 semivowels and 20 consonants) as compared to many thousands of different words. The standard method of ASR employs the basic principles of pattern recognition to discriminate among acoustic patterns. Creating a limited size of reference pattern is much easier than an infinite size of reference pattern. However, the problem of creating and storing an indefinite size of reference data can be resolved. On the other hand, the time required to process and compare the unknown data with the reference data of a limited size is much less than that required for an unlimited size of data.

In the past fifteen years, extensive study on phoneme recognition has been done by a number of ASR researchers[7]-[30]. Typical phonemes that they have studied are English stop consonants /p,t,k,b,d,g/. The reason is that the stop consonants occur very often in the English language[31]. At present, many models using acoustic features such as transitional cues, duration, silence and voicing, have been developed to recognize stops. But most of them require human assistance and are only limited to isolated words. In this research, the author proposes an automatic speaker-independent system using three distinctive acoustic features to recognize English stops in continuous speech went into limbo because many difficulties were encountered. It was also found that continuous speech recognition was much harder than isolated word recognition. As a result a large number of continuous speech projects were dropped. Somehow there was a turning point in the early seventies. The Defense Advanced Research Projects Agency (DARPA) foresaw the future potential of continuous speech recognition (CSR). They suddenly put millions of dollars annually into speech understanding research of which a substantial portion was allocated to CSR, which spurred the current interest in the problem. Meanwhile, big computer companies such as IBM, NEC and Sperry-Univac also started launching sizeable CSR projects.

I.2 MOTIVATION

As mentioned earlier, ASR provides a very important link in man-machine communication. It can minimize a tremendous amount of paper-work such as the key-punching of data information. It can resolve the problems facing those handicaps who cannot write or hear. An ideal system can receive speech signals and interpret the contents into different representations which can be recognized by anybody. For example, the information can be translated into printed materials which can be seen by the deaf. As well, people can directly talk to the ASR system to update the inventory information in commercial applications.

Table I-1. Summary of Stop Consonants Recognition and Related Experiments

Author	Date	Features Used	Features Extr. By	Decision Method	Materials Used	Recognition
Tang	1981	FT,SI & VOT	Automatic	Min. Dist.	Unconstrainted continuous speech spoken by 3 male 3 female speakers (120 sentences)	79.58% for stressed data (48 sentences). & 70.54% for unstressed data (72 sentences)
Santerre & Suen	1981	FT,SI,VD & VOT	Manual	NA	Isolated words spo-ken by 3 male & 3 female speakers	NA
Blumstein & Stevens	1980	VD	Manual	Human perception	Synthesized initial stops in isolated syllables	Over 80%
Datta et.al	1980	FT & VOT combined	Manual	Statistical techniques with para-metric rep-resentation	Isolated words by 3 male speakers (600 CV words)	Max. 74.9%
Demichelis et.al (Proposal)	1979	FT & VOT combined	Automatic	Fuzzy algorithm	Pseudo-syllables for voiced & CV syllables for unvoiced stops in continuous speech	NA
Port	1979	SI	Manual	Human perception	Syllables in running speech	NA
Pal & Majumder	1978	VOT & FT	Manual	"	PB CVC context spoken by 5 male speakers (600 samples)	60% for dentals & 85% for bilabials

Searle et.al	1978	VOT spectral peaks, shapes	Filter bank	Discrimin. analysis	Initial stops in isolated words (148 words)	77%
Wolf	1978	FT, SI, VD & VOT	Manual	Human perception	Six repetitions syllables spoken by two male speakers	~68.72% combined
Lisker et.al	1977	VOT	Manual	Manual	Synthesized CV syllables	NA
Molho	1976	Autocorrelation coef.	Automatic	Fuzzy algorithm	21 sentences spoken by 4 speakers	~50% for "p,t,k" & 66.7% for "d"
Itakura	1975	"	"	Minimum prediction residual	Designated male speaker via telephone input in isolated words (200 Japanese words)	97.3%
Weinstein et.al	1975	FT & VOT	Manual	Ebst & Fbst analysis	Segmented phonemes in connected speech	76% for d,g,t,k
Winitiz et.al	1975	VOT	Manual	NA	CV monosyllables spoken by one male speaker	NA
Cole & Scott	1974	FT & VOT	Manual	NA	Tape spliced initial stop syllables	Average 93.83% & 75.83% with target vowels i,u respec.
Suen & Beddoes	1974	SI	Manual	NA	Word pairs spoken by 3 male & 3 female speakers	NA
Stevens & Klatt	1974	FT & VOT	Manual	Human perception	Synthetic CV syllables	NA
Eimas & Corbit	1973	VOT	Manual	Human perception	Synthetic speech	NA

without human intervention: viz., Formant Transitions (F.T.), Silent Interval (S.I.), Voice-Onset-Time (V.O.T.).

Table I is the summary of stop consonants' recognition and related experiments from other researchers. The partial result of the present research is also included.

I.3 GOALS OF THE THESIS

The basic aim of this research is to propose an automatic system to recognize English stop consonants in continuous speech. In this research, the author would like to pursue the following objectives:

a). Automation - the system is automatic in the sense that there is no human intervention required.

b). Adaptability - the system can adapt itself to recognize any speakers for any length of speech.

c). Efficiency - the system should be efficient enough to minimize time and storage.

d). Modification - the system can be easily modified to recognize other phonemes.

I.4 CSRS SCHEME

Figure I-1 shows a diagram of the proposed continuous speech recognition system (CSRS). The function of this system is to detect and extract the common features from the input

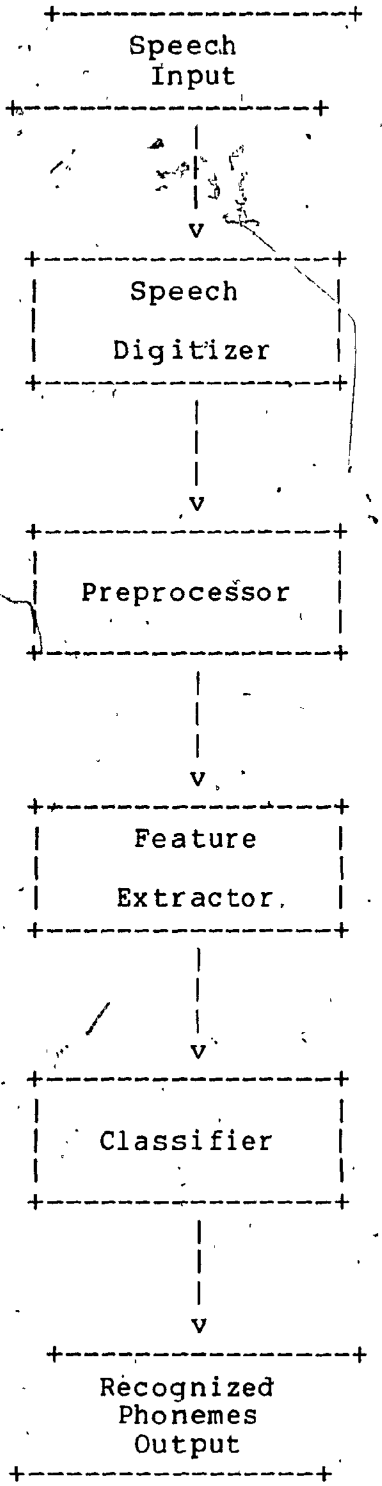


Figure I-1. Basic Structure of the Proposed Continuous Speech Recognition System

speech signals, and to use these features to recognize and classify the signals into different phoneme classes. The system includes four main stages - digitization, preprocessing, feature extraction and classification. All signals must pass through these four stages before they can be identified.

In the first stage, the input speech signals are quantized into digital data for computer processing by an 8-bit A/D converter. After the signals have been quantized, they are sent to the second stage for preprocessing, where a 12.8-msec segment of waveform is selected at one time and transformed into a frequency spectrum for further processing. Then the first two formants are extracted from the frequency spectra in order to make them ready for formant measurement. A smoothing teehnique is applied at this stage to obtain smooth formants. The advantage is to minimize the errors which may result from later processing. In the next stage, the extracted formants are tested by three different methods - F.T. (Formant Transitions), S.I. (Silent Interval) and V.O.T. (Voice-Onset-Time). The features are extracted from the formants based on the measurements made in each method. After the features have been extracted, they come to the final stage - the classification stage - for identification. Each type of the derived features is examined to determine the identification of the input signals. There are two phases in this stage. The first phase is called the training phase, in which each classifier is trained before usage. The second phase is the decision phase. The input phoneme is classified into the most acceptable class based on the distance measurements.

The whole system, with the exception of the digitization system which is implemented in the INTEL 8085 micro computer, is programed in the Fortran IV computer language and processed by a CDC Cyber-172 computer.

I.5 OVERVIEW OF THE THESIS

This thesis is composed of five chapters. Chapter I presents the historical background of the field of automatic speech recognition. Research and development of ASR in the past four decades are described. The following sections discuss the motivation of this study and the goals to be achieved in this work. The basic structure of the proposed CSR system and the contents of this thesis are outlined at the end of this chapter.

The steps involved in the database setup, the components of the data collection system, the devices and methods used to convert the input signal into a form suitable for the recognition system are discussed in detail in Chapter II.

Chapter III describes the organization and the functions of CSRS. Three different methods are used to detect and extract the features from the input data to train the classifiers. The minimum distance measure technique was used to classify the phonemes based on the characteristics of the extracted features of each method.

Experiments conducted in this research using the proposed CSRS are discussed in Chapter 4. Statistical results of each method are presented at the end of each major section.

Finally, Chapter V includes the whole CSR system review, the conclusions of the present research, and suggestions for further study.

CHAPTER II

DATABASE AND DATA PREPROCESSING

II.1 DATABASE

Data collected for this study include 160 unstressed and 160 stressed sentences uttered by eight untrained paid native speakers of English, four males and four females. They belong to the age group of twenty to forty. The following section discusses this experiment in detail.

II.1.1 Data Collection

The recordings of sentences were made in two sessions in a 12ft x 15ft x 10ft sound-proof room on two 1200-ft Scotch tapes at the speed of 7 1/2 ips. Figure II-1 shows the recording system setup which consists of one Sennheiser MD 421U dynamic cardiod microphone, one Tascam Model 10 mixer, one Ampex AG 440B tape recorder and one signal light.

During the recording, one page of computer-printed material which contained 20 sentences of medium size, was given to each speaker. The speakers were instructed, one at a time, to sit alone in front of the microphone in the sound-proof room. Depending on the loudness of the voice of the speaker, the distance of the microphone from the speaker and the mixer volume were adjusted to produce a preset output level. A green light box placed in front of the speaker was

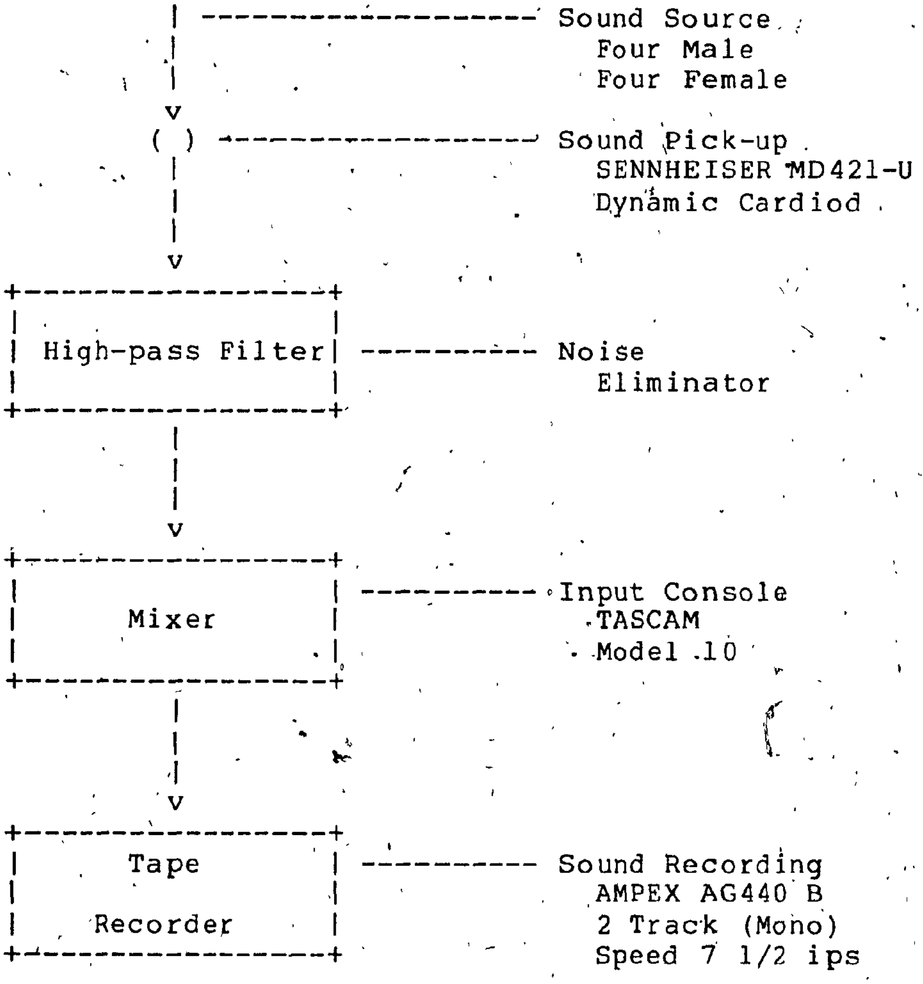


Figure II-1. Recording System Setup

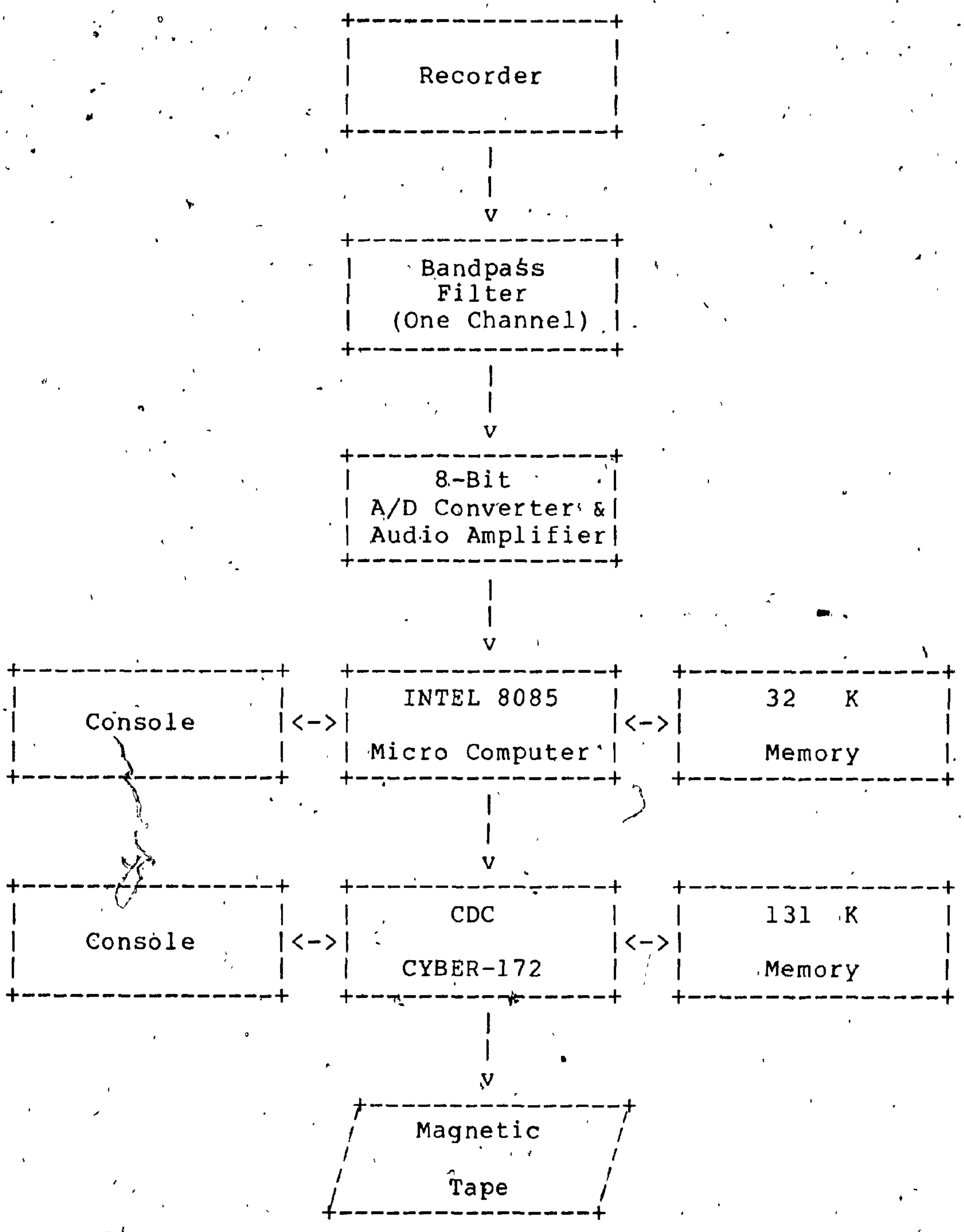

Recorder
Bandpass Filter (One Channel)
8-Bit A/D Converter & Audio Amplifier
Console
INTEL 8085 Micro Computer
32 K Memory
Console
CDC CYBER-172
131 K Memory
Magnetic Tape

Figure II-2. Speech Digitization System

used to signal him when to start reading the sentence. Before actual recording took place, the speakers were asked to read several test sentences according to the signal given until they were familiar with the system. Each speaker was given a three-second time interval between two sentences in order to prepare him for the next sentence.

The recording was divided into two parts. In the first part, the speakers were asked to read the sentences in their usual way. In the second part, the speakers were required to read the same sentences again with stress on those stop consonants indicated by arrows (see APPENDIX I). The purpose was to ensure that the speakers had not missed enunciating any stop consonants.

In this experiment, there were no constraints, such as speech speed, intonation etc., imposed on any speakers. Therefore, artificial speech was avoided. Altogether 320 sentences which contained 2400 stop consonants were collected.

II.1.2 Speech Digitization

After the speech samples have been recorded, they are ready for the next processing step. Since the computer can only process digital information, speech has to be converted into digital signals. Figure II-2 shows the schematic diagram of the speech digitization system. It consists of one Sony 2-track mono tape recorder, one band-pass filter, one 8-bit

(256 levels) Analog/Digital (A/D) converter, one INTEL 8085 micro computer and one CDC Cyber-172 computer.

Initially, the speech tape is played back on a Sony tape recorder at the same speed as recorded (7 1/2 ips). The speech signal passed through a 200 Hz - 9 kHz bandpass filter so that frequency components outside this range is filtered out at -10 dB attenuation. This should minimize the noise created by machines, electronic devices, or other sources. The filtered signal is amplified to 4 volts before entering the A/D converter. A strong signal may resolve the problem of distinguishing weak speech from silent speech. The function of the A/D converter is to convert the audio signal to digital form. It is directly connected to the INTEL 8085 micro computer. It converts the signal into 256 levels at a sampling rate of 10 kHz which can provide enough information for subsequent processing (Markel[6]). Due to the limited memory size of the micro computer, only 3.2 seconds of speech signal can be processed at one time. If the speech signal is longer than this, it is split into two or more segments. Once the speech signal has been digitized, it is transferred to the CDC Cyber-172 computer for storage. All digitized signals are stored on two 2400-ft 1600 BPI magnetic tapes.

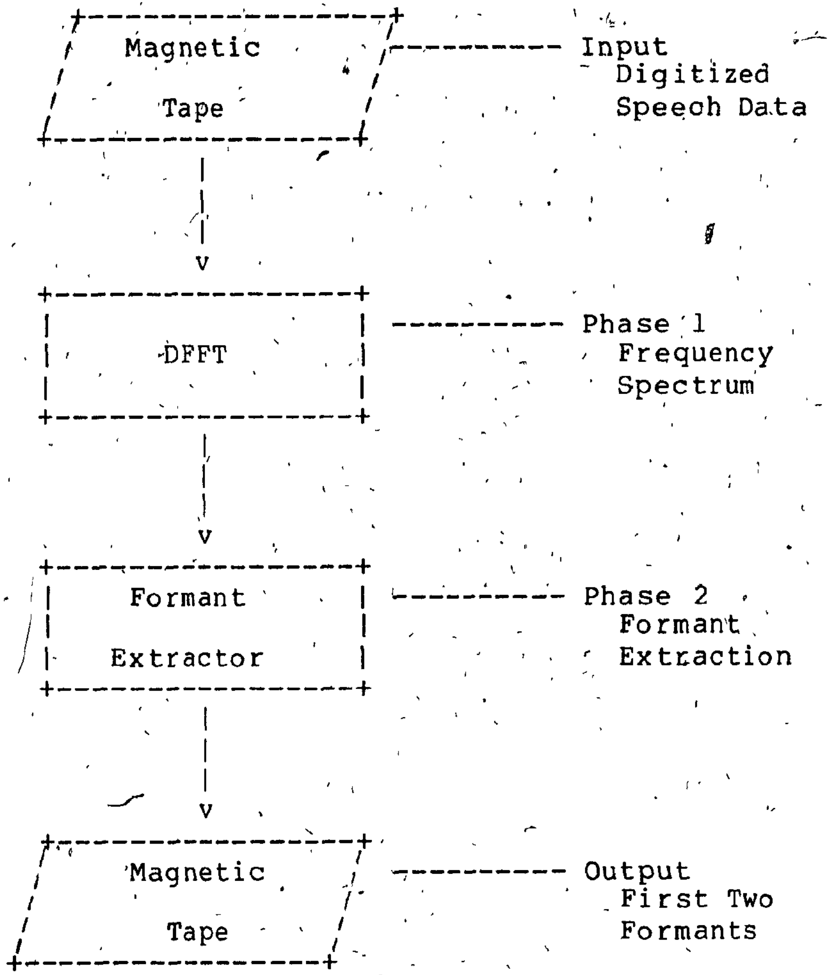


Figure II-3. Data Preprocessing System

II.2 DATA PREPROCESSING

Data preprocessing is one of the most important steps in speech processing. The ideal preprocessing algorithm will achieve higher recognition scores and minimize unnecessary processing time and storage. Two different approaches have been developed and applied by a number of researchers to preprocess the input data. One approach is hardware oriented. Special electronic circuits (Flanagan [32]) or expensive sono-graphic machines to produce speech formants are widely used. Another approach, which is more economical and can fully utilize the existing computer system, is software oriented. Bergland[33], Markel[34], Oppenheim[35] and Schafer and Rabiner[36] suggested the use of Fourier series to obtain speech spectra. Markel[6], McCandless[37] and Schaefer and Rabiner[36] also suggested the use of peak picking method to obtain speech formants.

In this research, the author employs the economical approach - software oriented to obtain the first two formants. Usually the variations in the first two formants of stop consonants are more significant than the higher formants. Therefore, the first two formants should provide enough information for formant analysis. The details will be described. In this system, the preprocessing stage contains two phases (Figure II-3), speech spectrum and formant extraction.

II.2.1 Speech Spectrum

The purpose of this first phase of preprocessing is to convert the speech waveform from the time-domain into the frequency-domain (Figure II-4a-b). The method employed is called Discrete Fast Fourier Transform (DFFT). It identifies the frequency components at each segment of the coming waveform. Time segment is set at 12.8 msec. At a sampling rate of 10 kHz, a total of 128 points are considered each time. As pointed out in the previous section, the first two formants are usually located at a frequency range between 100 Hz and 5 kHz. Hence a 10 kHz sampling rate provides enough information for a spectral analysis of the speech signal.

The following formulae are used to compute the DFFT, frequency resolution and frequency range after DFFT:

DFFT formula:

$$S(f) = 1/n \sum_{t=0}^{n-1} x(t) \cdot e^{-i2\pi ft/n}$$

where $x(t)$ = time-domain function;

n = no. of discrete points;

$i = \sqrt{-1}$.

Frequency resolution (df):

$df = 1/dt$

where dt = time resolution.

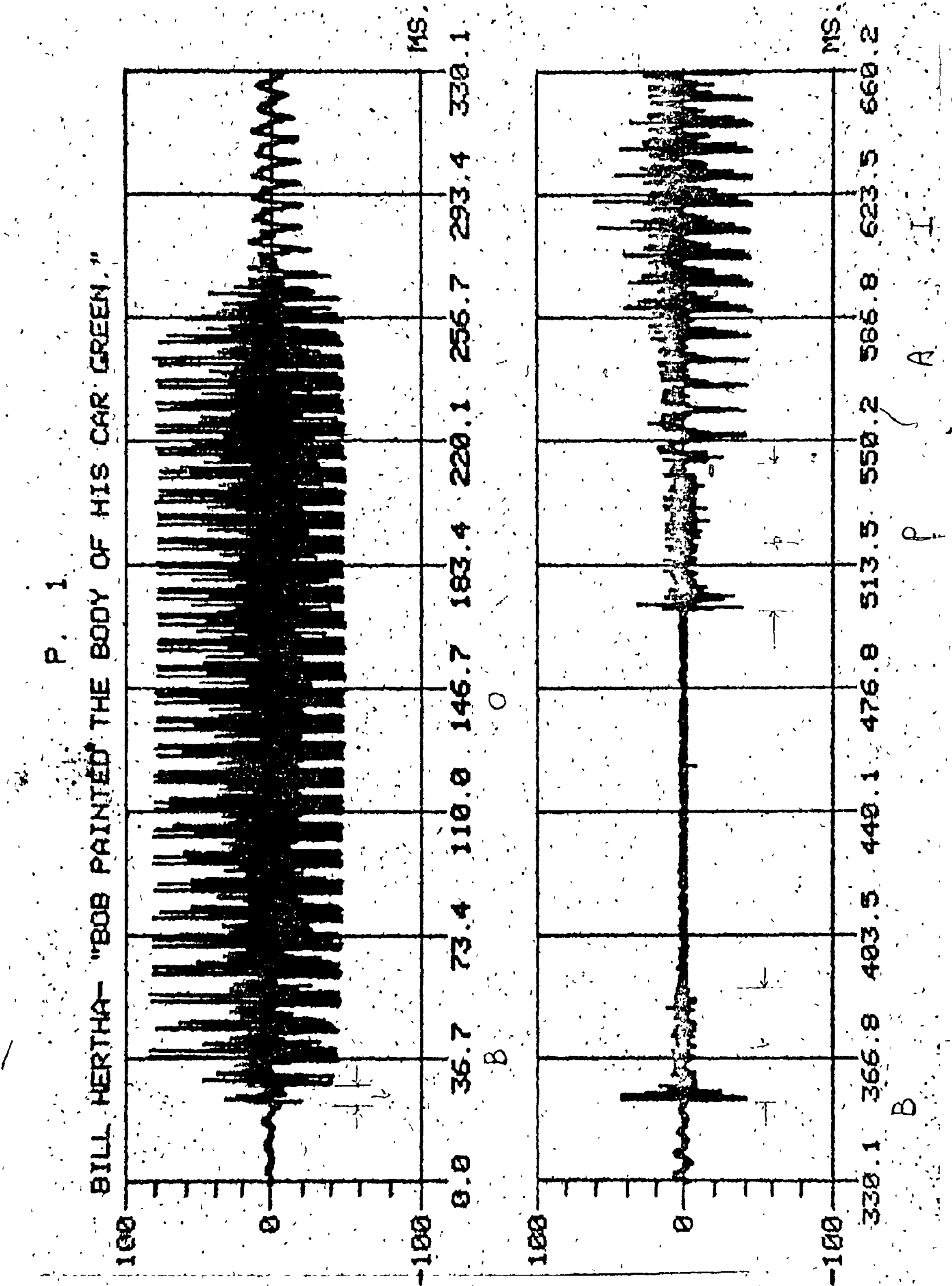


Figure II-4a. Speech Signal of the word "Bob" After A/D Converter

```
         DATE - 81/08/04.     TIME - 17.48.49.

           BILL HERTHA- "BOB PAINTED THE BODY OF HIS CAR GREEN."

      0.00   .39  .78 1.17 1.56 1.95 2.34 2.73 3.13 3.52 (KHZ)
      0.  |----|----|----|----|----|----|----|----|----|---->
  12.80 |
  25.60 | BAAAAAAAAAA
  38.40 |BBBBCECGBCFAGABAAABAAB A A A   AABAA A   AAAA
  51.20 |BBBBBIDICDAHBGABAAA A  A A     AABCAAAAAAA AAAA A
  64.00 +BBBACHCJDCEIHECBBAAAABBAA ABAAA  BBBB AAA  A AAA
  76.80 |B CAAEBJDCENGEBCAABBABAAAAAAAAAABACABAAAAAA A  A A
  89.60 |ABCACFAJDBJGMDBCCACC A BB BBAB ABADDBBB BAA  AA A
 102.40 |AABACDDHCCJHNBB AAABBAABAAAABABAB DAC AAA    AAAA
 115.20 |ABB DEEKDBMASDC ACDABAACAAAABABBABDBDAAAAAA AAABAA
 128.00 +ABBADDDHCGKFPAABAAABCABBABA BABABADBBAABB AAAAABAA
 140.80 |AABABDDGCIKDR  A BAACAAABAAABABAAAADD BAAAAAA AABA
 153.60 | BCADCGKAEDIRCBB   ABBB CBAAABC AAAD BBACAAAAAAAAA
 166.40 | ACADBGJFFEFKJDCCBBBBCAABA  BBBBAAABCAABABAAAABABB
 179.20 |AABAACCGIHCHJMCCCBBCABAABABAAABAACABDAAAAAAA  AAAA
 192.00 +ABBAB EDOEDKCPABAAB B BAABAAAAAABADCDABB BAAAAAAAA
 204.80 | BBAADGAOBDEKGIEBCBBAAAABAA BAB BDADBDAAABBAA AABA
 217.60 |BCBBAFHEMEHAKBKBBBAAABAAABABAAAABCCCCAABAAB BBAA A
 230.40 |BAABCEGICMLCLCICCAAAB BAAAB A AAACBGABAAAAABAAABAA
 243.20 |BAABBDFMCKICJHFFBCBACAB A B A AAAADCDBBBABABAAAB A
 256.00 +BECEEBHIADDAJHGDACAB A AAAAA A AACEDCABAA AA AAAAA
 268.80 |BCCAFCAB ABBFDCAAAAAA A A AAAAAAAECBAAAA     A AAA
 281.60 |AEDBBA A   ABAA                AAAA A
 294.40 | E B A       A                  A
 307.20 | E B A
 320.00 + EAB
 332.80 | C A
 345.60 | B
 358.40 |AAAAA   AAAA    AABBBBAAAAAABBBBBBAAA AAAAA  AA
 371.20 |A  A              AA A    AA    AA
 384.00 + A                    A    A     A              A
 396.80 |
 409.60 |
 422.40 |
 435.20 |
 448.00 +
     .    .      A .. Z, 0 .. 9 - Amplitude levels.
     .    .          where A represents lowest level;
     .    .                9 represents highest level.
          v
```

Figure II-4b. Speech Spectrum of the word "Bob" After DFFT

II.2.2 Formant Extraction

The present study is mainly based on formant analysis. Therefore a greater effort has been put into it to develop an efficient formant extraction algorithm.

Formant extraction forms the second phase in the preprocessing stage. In this thesis, an automatic formant extraction algorithm has been developed to extract the first two formants from the speech spectrum. The algorithm is based on peak picking. An interpolation technique has been applied in order to obtain smooth formants.

II.2.2.1 Peak Picking

The first two formants of the speech spectrum are extracted through peak picking. It simply selects the first two peaks in the spectrum and calls them the first two formants. A peak is determined by selecting the point at the highest amplitude within the specified frequency range of the spectrum. Figure II-5 shows the first two formants extracted from a speech spectrum.

The results of this experiment show that the first formant (F1) usually falls between 100 Hz to 850 Hz while the second formant (F2) usually falls between 550 Hz to 2500 Hz for both male and female voices. On the other hand, they also show that the formant extraction algorithm works very well except

```
        DATE - 81/08/04.     TIME - 17.48.49.

        BILL HERTHA- "BOB PAINTED THE BODY OF HIS CAR GREEN."

     0.00  .39  .78 1.17 1.56 1.95 2.34 2.73 3.13 3.52 (KHZ)
     0. |----|----|----|----|----|----|----|----|----|---->
  12.80 |
  25.60 | .        "
  38.40 |       .    "
  51.20 |      .    "
  64.00 +       .   "
  76.80 |       .   "
  89.60 |       .     "
 102.40 |       .     "
 115.20 |       .     "
 128.00 +       .     "
 140.80 |         .          "
 153.60 |       .     "
 166.40 |       .     "
 179.20 |        .     "
 192.00 +      .     "
 204.80 |        .    "
 217.60 |        .     "
 230.40 |         .     "
 243.20 |       .    "
 256.00 +       .    "
 268.80 |    .       "
 281.60 | .          "
 294.40 | .            "
 307.20 | .
 320.00 + .
 332.80 | .
 345.60 | .
 358.40 |   .                 "
 371.20 |   .                        "
 384.00 + .                         "
 396.80 |
 409.60 |
 422.40 |
 435.20 |
 448.00 +
    .   .     where . represents the first formant;
    .   .           " represents the second formant.
    .   .
        V
```

Figure II-5. First and Second Formants of the word "Bob" Extracted by Formant Extraction Algorithm

for the following situations: a) Two peaks of equal amplitude show up, the decision to pick up the right one may be erroneous; b) Occasionally noise created by the speaker (such as heavy breathing) or by the machine (such as the machine control button switched on and off) may be incorrectly interpreted as part of a formant; c) Segmentation for each formant frequency range may not be appropriate to every word, thus a wrong peak may be picked up. As an example, the peak at the time frame 140.80 msec in Figure II-5 is out of line due to improper formant frequence segmentation.

II.2.2.2 Formant Smoothing

Since formants reflect the movements of the vocal tract, they can change considerably within a short period, e.g. at the boundary between a nasal and a vowel. Therefore, rough formants may be obtained. In order to obtain smooth formants and to solve the problem of picking the wrong peaks, a formant smoothing algorithm based on an interpolation technique has been applied. It works as follows: a) If a formant is missing in the spectrum, fill in its frequency with an average value of the previous and the following formant frames; b) If a formant is out of line, correct it by interpolation as follows:

Let $D_{m,n}$, F and THR be the difference between two formant frames, the formant and the threshold respectively. If $D_{n,n-1}$

```
DATE - 81/08/04.     TIME - 17.48.49.

BILL HERTHA- "BOB PAINTED THE BODY OF HIS CAR GREEN."

     0.00  .39  .78 1.17 1.56 1.95 2.34 2.73 3.13 3.52 (KHZ)
     0. |----|----|----|----|----|----|----|----|----|---->
  12.80 |
  25.60 |      .   "
  38.40 |      .   "
  51.20 |      .   "
  64.00 +      .   "
  76.80 |      .    "
  89.60 |      .    "
 102.40 |      .    "
 115.20 |      .    "
 128.00 +       .   "
 140.80 |       .   "
 153.60 |       .   "
 166.40 |       .    "
 179.20 |       .    "
 192.00 +       .    "
 204.80 |       .    "
 217.60 |       .    "
 230.40 |       .    "
 243.20 |       .    "
 256.00 +      .     "
 268.80 |     .      "
 281.60 |   .        "
 294.40 |  .         "
 307.20 | .
 320.00 + .
 332.80 | .
 345.60 | .
 358.40 |   .                      "
 371.20 |  .                       "
 384.00 + .                        "
 396.80 |
 409.60 |
 422.40 |
 435.20 |
 448.00 +
    .   .    where . represents the first formant;
    .   .          " represents the second formant.
    .   .
        V
```

Figure II-6. First and Second Formants of the word "Bob" Smoothed By Interpolation

< THR, i.e. the current formant is smooth, then skip to the next frame; otherwise, perform one of the following operations:

1) If $D_{n+1,n-1}$ < THR and $D_{n+2,n+1}$ < THR, then
$F_n = (F_{n+1} + F_{n-1}) / 2$

2) If $D_{n+2,n-1}$ < THR and $D_{n+3,n+2}$ < THR, then
$F_n = (F_{n+2} + F_{n-1}) / 2$

3) If $D_{n+3,n-1}$ < THR and $D_{n+4,n+3}$ < THR, then
$F_n = (F_{n+3} + F_{n-1}) / 2$

and smooth the current formant twice using the following formula:

$$F_n = 1/4\ F_{n-1} + 1/2\ F_n + 1/4\ F_{n+1}$$

Figure II-6 shows the first two formants after smoothing. A typical example showing the difference between formants before and after smoothing with a time frame of 140.8 msec can be found in Figure II-5 and in Figure II-6. In the former figure, the value of the second formant is far away from the previous and the following formants. But after smoothing, the formant has been aligned with the formant track.

CHAPTER III

RECOGNITION ALGORITHMS AND FEATURES USED

In this chapter, three different methods have been developed to recognize /p,t,k/ and /b,d,g/. They are 1) Formant Transitions (F.T.), 2) Silent Interval (S.I.) and 3) Voice-Onset-Time (V.O.T.). The classification technique for each method will also be discussed at the end of each section. The organization of the recognition system is shown in figure III-1. This system, in fact, contains three subsystems. Each of them works independently. The values of the first two formants are sent to these subsystems, from which each subsystem will extract its own features. Subsequently, the features pass through a classifier to be separated into different categories of stop consonants. Each subsystem has been designed to process one sentence at a time.

III.1 FORMANT TRANSITIONS

As pointed out by Cole and Scott[12], Datta, Ganguli and Ray[13], Menon, Rao and Thosar[14], Pal and Majumder[15], Santerre and Suen[16], Sharf and Hemyer[17], and Wolf[18] there is a rapid change in the vocal tract shape which makes the transition from one place of articulation to another when a stop consonant is uttered with a preceding or following vowel. They also conclude that the change in formant frequency (transitional cue) of the vowel associated with the stop consonant(s) may provide information general enough to distinguish voiced and unvoiced stops in most cases. This method simply computes the percentage change of the first two formants from the plosive release to the steady state of the associated vowel for preceding stops (including initial and medial stops), and from the steady state to the closure for final stops. Here, the plosive release is defined as the burst period when a stop is uttered; the steady state is the formant frequency which fluctuates less in the voicing period; and the closure is the silent period which the vocal tract is closed to prepare the following burst to release. Figure III-2 shows the measurements of the first two formants for preceding and final stops. F1 and F2 represent the first and

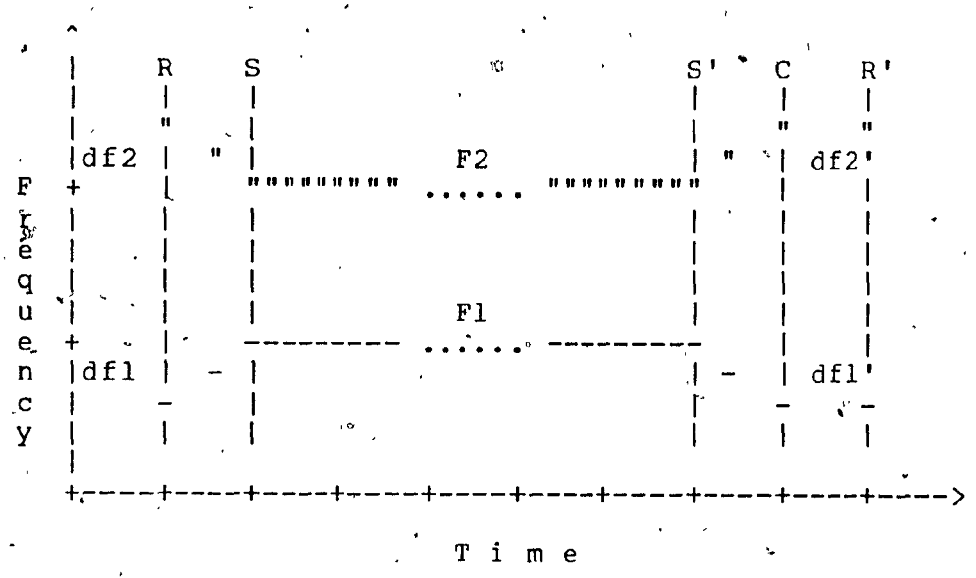


Figure III-2. Measurements of Formant Transition Frequencies.

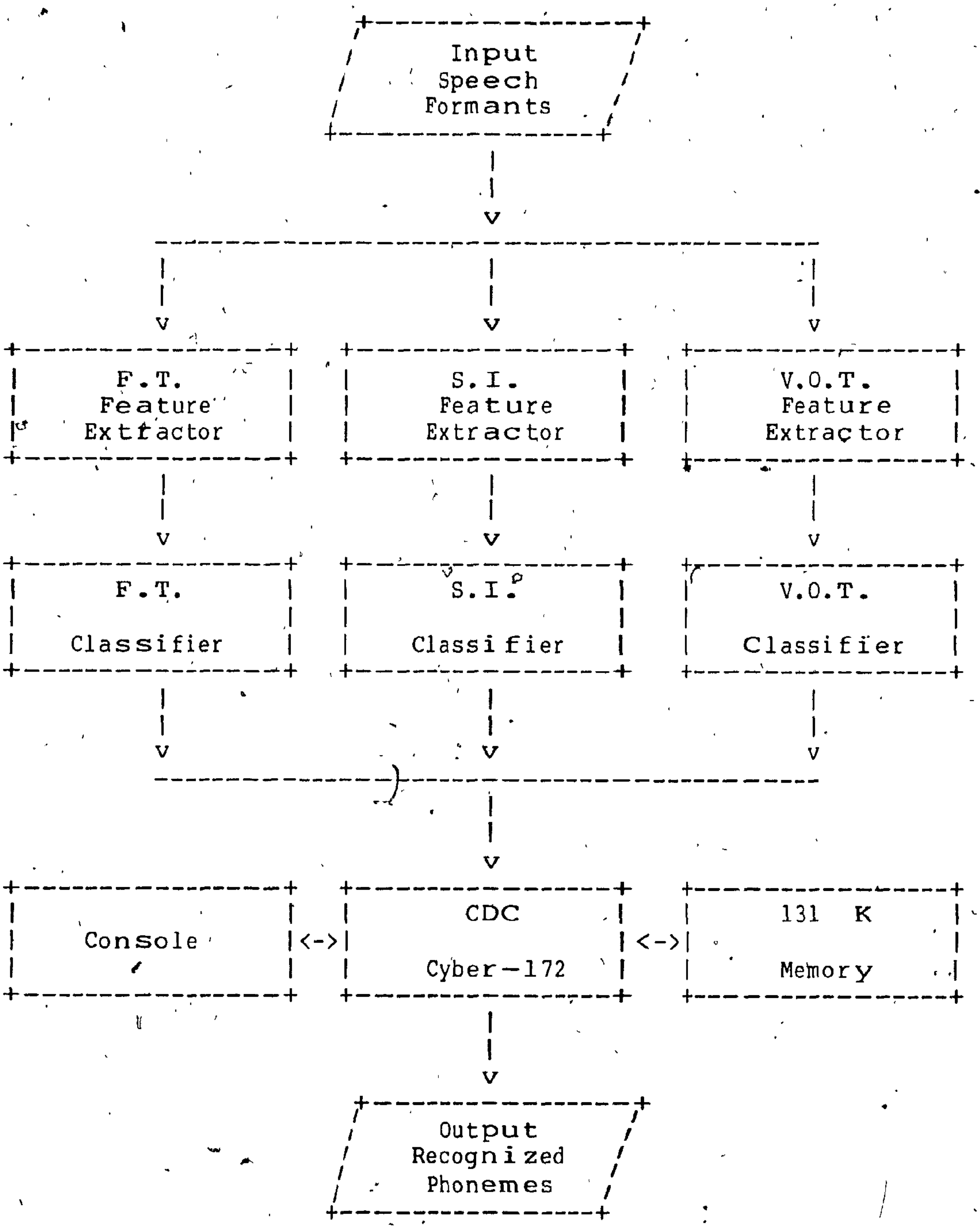


Figure III-1. The Organization of the Stop Consonants Recognition System

the second formants, df1 and df2 represent the change in the frequency of preceding stop from R to S of F1 and F2, and df1' and df2' represent the change in frequency of the final stop from S' to C of F1 and F2 respectively. The percentage changes of transition are computed as follows:

Preceding Stops -

Let FT1 and FT2 be the percentage frequency changes of a preceding stop in F1 and F2, and f1r and f2r be the frequencies of F1 and F2 at plosive release starting point (R).

$$FT1 = (df1 / f1r) \times 100\% \quad (3.1)$$

$$FT2 = (df2 / f2r) \times 100\% \quad (3.2)$$

Final Stops -

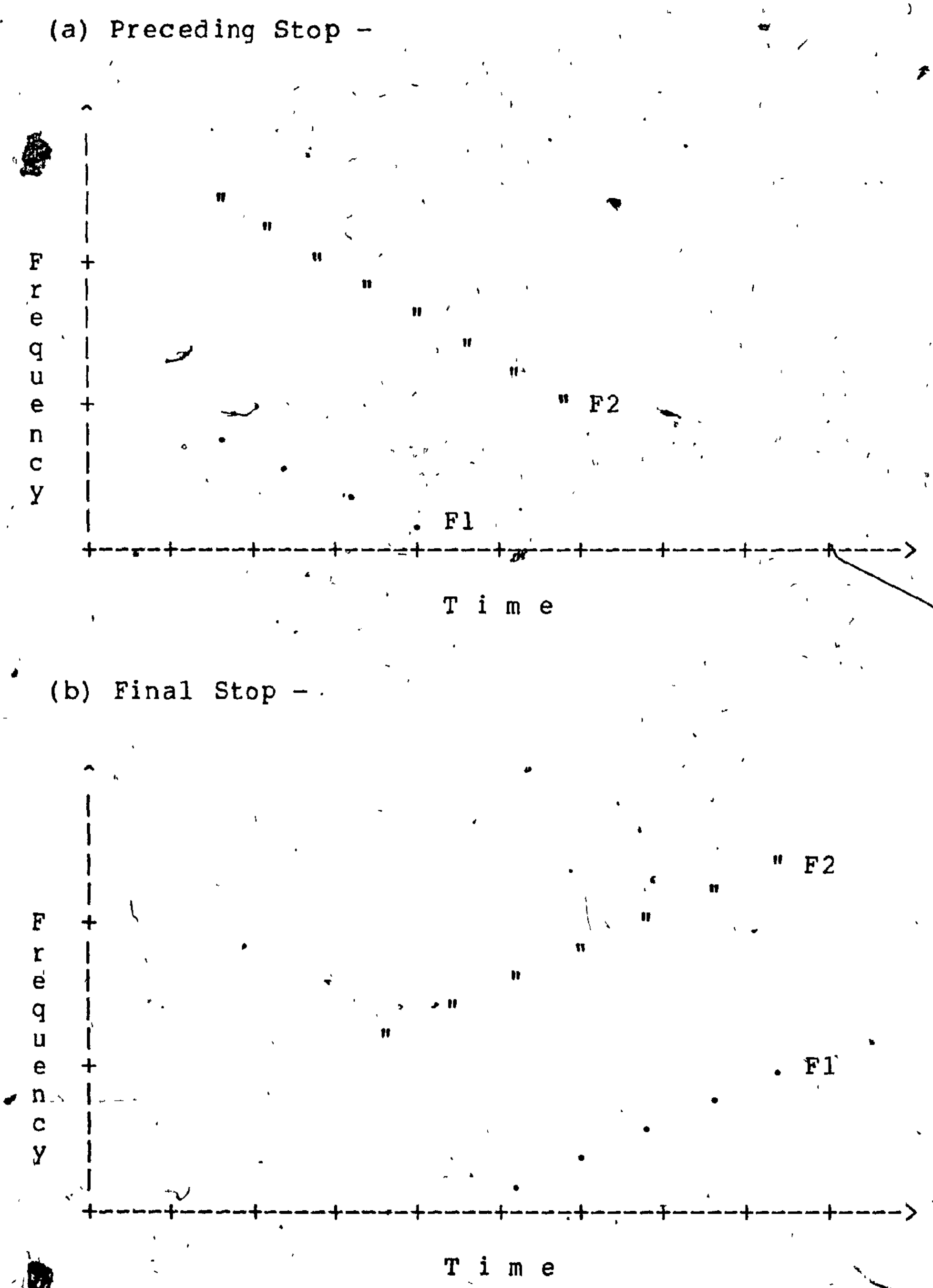
Let FT1' and FT2' be the percentage change of a final stop in F1 and F2, and f1s and f2s be the frequencies of F1 and F2 at steady state.

$$FT1' = (df1' / f1s) \times 100\% \quad (3.3)$$

$$FT2' = (df2' / f2s) \times 100\% \quad (3.4)$$

III.1.1 Feature Extraction

Feature extraction is the most important step in the formant transition classification process. The outcome of each classified phoneme is dependent on this procedure. In the present research a fully automatic F.T. feature extraction



(a) Preceding Stop -

(b) Final Stop -

Figure III-3. Steady state does not apparently occur in (a) Preceding Stop, and (b) Final Stop

algorithm is introduced.

The procedure of this algorithm always searches for the preceding stop first, and then the final stop. The algorithm initially determines the starting point (R) which is the place of articulation of the preceding stop. Usually, there is a pause before the burst releases. The algorithm searches for it to determine the starting point. Once the algorithm has detected this point, it will search for the steady state (S) (Figure III-2). The steady state is confirmed if three or more consecutive frames have about the same formant frequency values. If it has been found, then the algorithm will compute FT1 and FT2 using equations 3.1 and 3.2; otherwise, the algorithm may assume the stop consonant does not exist in the syllable and will proceed to look for the final stop.

According to the characteristics of final stops, a stop exists if there is a burst after the steady state (S'). However, there is a closure period called the silent interval which will be discussed in the later section, between the closure starting point and plosive release starting point, i.e. the period between C and R'. Therefore, in order to find the final stop, the algorithm will search for the steady state first. If it is found, then the algorithm will proceed to search for the closure starting point, and compute FT1' and FT2' using equations 3.3 and 3.4; otherwise, it will go back to the first step to repeat the same procedure until the whole sentence is finished.

The results show that the algorithm works quite satisfactorily except in the following cases: in Figure III-3a-b, stop consonants actually exist in both cases. Due to substantial changes in frequency, the steady state cannot be found. Furthermore, if the word boundary or the pause is not clear, the starting point cannot be detected.

III.1.2 F.T. Classifier

Classification is the last stage in CSRS. The unknown speech input will be classified into the most acceptable class at this stage. However, if the feature does not satisfy the given conditions, it will be rejected. Before the classifier is used practically, it has to be trained. Therefore the classifier usually contains two phases: training and classification. The training phase, is merely a temporary phase, which is no longer used once the classifier has been trained. On the contrary, the classification phase is used permanently to recognize future features until modification is required in order to recognize other features. A typical pattern of the proposed speech recognition classifier is shown in Figure III-4.

In the training phase, the method used to train the classifier is based on a statistical approach. The steps involved are as follows: initially, the feature vectors which have the common characteristics are selected from the training

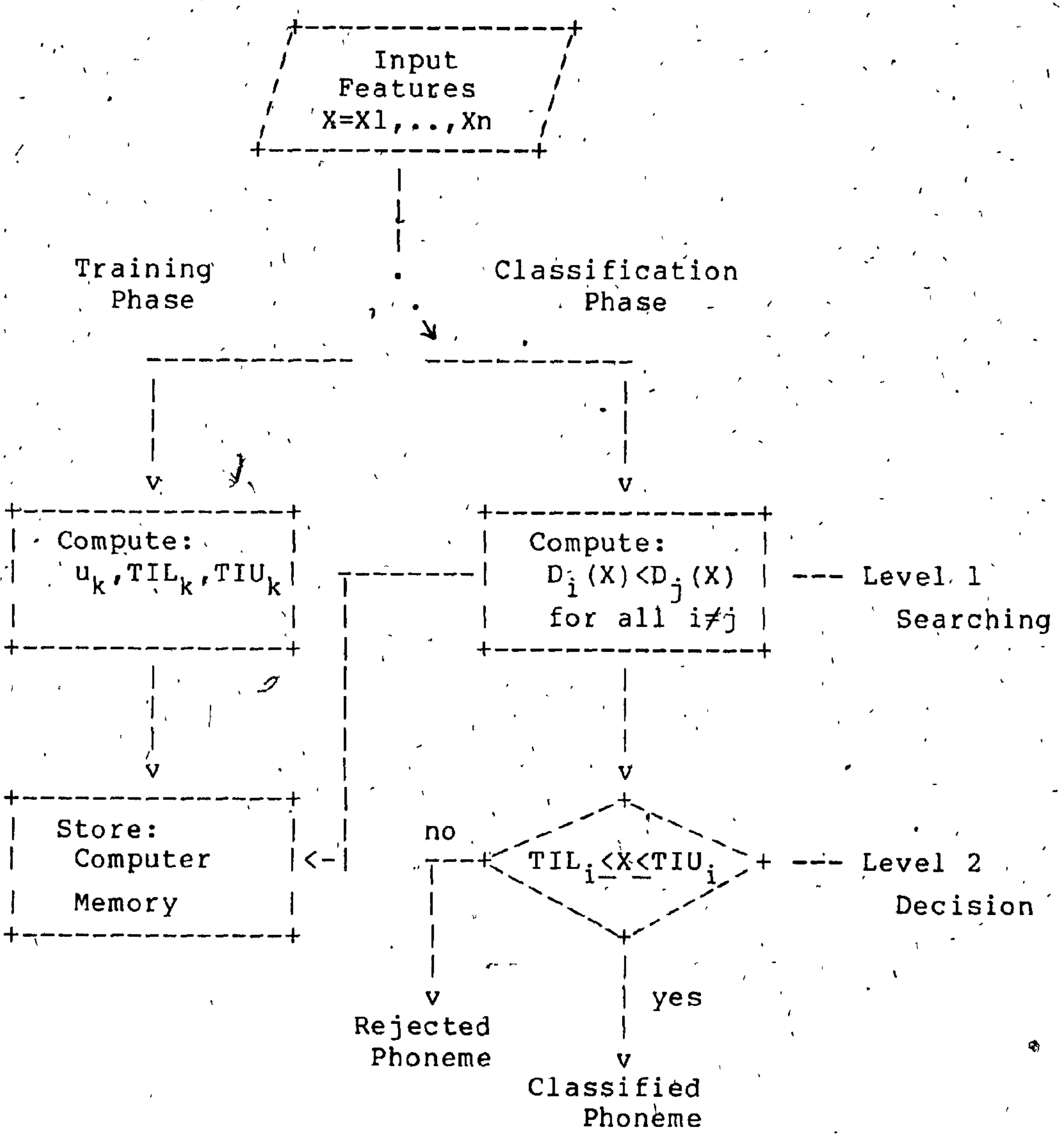


Figure III-4. A Typical Speech Recognition Classifier

data and put into different groups of stop consonants. There are six main groups of preceding stops and final stops respectively. Each group represents one stop and it is subdivided into thirteen subgroups or classes. Each of them represents a vowel labelled according to the International Phonetic Alphabet (IPA). In other words, each stop may have thirteen different combinations of vowels following the preceding stops, and thirteen different combinations of vowels preceding the final stops. The combinations of stop consonants associated with different vowels can be referred to in APPENDIX II. Altogether, two discriminant matrices are formed, one for the preceding stops and the other for the final stops (which contain 78 classes each). However, only 35 classes are used in the preceding stops classification and 24 classes are used in the final stops classification, because the other combinations do not exist in the data. In each class, it contains two types of features. The first one (FT1 or FT1') is extracted from the first formant, and the second one (FT2 or FT2') is from the second formant. The algorithm selects the feature vectors from each type to compute the mean and standard deviation using the following equations:

Mean (u) -

$$uFT_k = \frac{\sum_{i=1}^{N_k} FT_i}{N_k} \quad \text{------------------------} \quad (3.5)$$

Standard Deviation (s) -

$$sFT_k = \sqrt{\frac{FT_i^2 - N_k\ uFT_k^2}{N_k - 1}} \qquad (3.6)$$

where FT_i = Feature vector in class k;
k = No. of classes;
N = No. of samples in class k.

In the classification phase, it consists of two levels, Level 1 and Level 2. A set of discriminant functions $d_i(X)$ is considered for each input feature (X) (Duda and Hart [27]). If the input feature satisfies the following condition:

$$d_i(X) < d_j(X) \qquad \text{for all } j \neq i$$

i.e. if the i^{th} discriminant function, $d_i(X)$, has the smallest value for a feature X, then X will be sent to the second level. In this level, feature X will be examined. If it again satisfies the following condition:

$$TIL_i \leq X \leq TIU_i$$

where TIL_i = Lower limit of class i;
TIU_i = Upper limit of class i.

then X will be assigned to C_i. If not, it will be rejected. Therefore, the output of the classifier may contain four types of recognition results - correct recognition, misrecognition, rejection and misrejection.

The following equations (EQ. 3.7 and 3.8) are used to compute the discriminant scores based on distance measure:

Preceding Stops -

Let $uFT1_i$ and $uFT2_i$ be the means of percentage change in frequencies in F1 and F2 of a preceding stop of class i, and DFT_i be the distance between the means of $FT1_i$ and $FT2_i$ of a stored pattern and the input features X1 and X2.

$$DFT_i = \sqrt{((X1 - uFT1_i)^2 + (X2 - uFT2_i)^2)} \text{ -------- } (3.7)$$

where i = 1 .. 35.

Final Stops -

Let $uFT1'_i$ and $uFT2'_i$ be the means of percentage change in frequencies in F1 and F2 of a final stop of class i, and DFT'_i be the distance between the means of $FT1'_i$ and $FT2'_i$ of a stored pattern and the input features X1' and X2'.

$$DFT'_i = \sqrt{((X1' - uFT1'_i)^2 + (X2' - uFT2'_i)^2)} \text{ --- } (3.8)$$

where i = 1 .. 24.

After computing the score for each class, the algorithm will search for the class with the minimum distance. Once it has been found, it will be checked to see whether it satisfies the last condition - tolerance interval. Since the input patterns may contain non-stop consonants, the purpose of the checking is to reject this kind of pattern. As the number of features in each class is small, and the features are not normally distributed (by observation), the tolerance interval is computed by using Chebyshev's Inequality (E.Q. 3.9-1 and 3.10-1) [39].

Assume the input feature (X) as a random variable, the mean (u) and standard deviation (s) are known in each class. The probability of X that lies within the interval is dependent on the value of k which has been adjusted according to the speaker.

$$P\ (\ |X - u_i| \geq ks_i\ ) \leq 1 / k^2 \text{ ---------------- } (3.9)$$

$$P\ (\ |X - u_i| < ks_i\ ) \geq 1 - 1 / k^2 \text{ ----------- } (3.10)$$

where X = Input feature;

u_i = Mean of class i;

k = Constant;

s_i = Standard deviation of class i.

From above, we obtain the interval as follows:

Lower Bound -

$$TIL_i = u_i - ks_i \text{ -------------------------- } (3.9-1)$$

Upper Bound -

$$TIU_i = u_i + ks_i \text{ -------------------------- } (3.10-1)$$

Figure III-5 shows the accepted and rejected regions of class i. If the value of X lies beyond the acceptable region, the input feature X will be rejected.

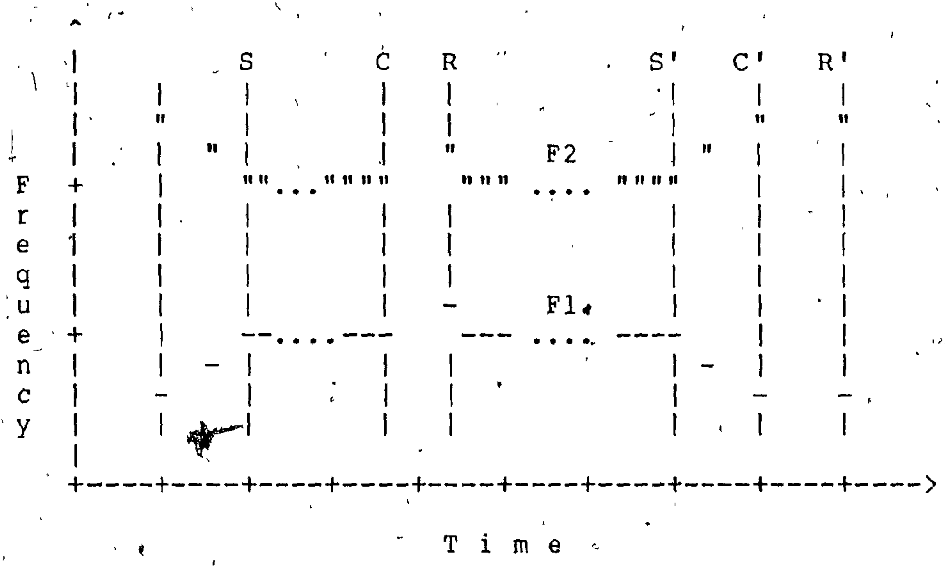


Figure III-6. Measurements of Silent Intervals.

together with a following vowel. It may be easily identified since there is a voicing after the burst has been released. Figure III-6 shows the measurements of S.I. for medial and final stops. SI1 and SI2 represent the time interval of the medial stop between the closure starting point (C) and the plosive release starting point (R) of F1 and F2 respectively. Similarly, SI1' and SI2' represent the time interval of the final stop between C' and R' of F1 and F2 respectively. The silent intervals are computed as follows:

Medial Stops -

$$SI1 = r1 - c1 \quad\text{------------------------------------}\quad (3.11)$$

$$SI2 = r2 - c2 \quad\text{------------------------------------}\quad (3.12)$$

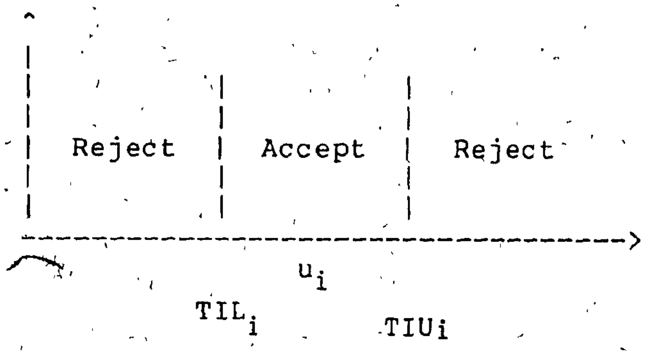


Figure III-5. Tolerance Interval

III.2 SILENT INTERVAL

As suggested by Cole and Scott[12], Liberman et.al.[19], Lisker[20], Port[21], Santerre and Suen[16], Slis et. al.[22][23], Suen et. al.[24], and Wolf[18], the S.I. is another important cue to distinguish voiced and unvoiced stop consonants. The S.I. is defined as the duration between closure and plosive release. The above authors conclude that voiced stop consonants usually have shorter duration of S.I. than unvoiced ones.

As noted in the previous method, there is a pause - silent period before the burst of a stop consonant. In this study, the S.I. is the second cue suggested to recognize medial and final stops. The medial stop is defined as the stop consonant which is located in the middle of the word. It is uttered

Final Stops -

$$SI1' = r1' - c1' \qquad (3.13)$$

$$SI2' = r2' - c2' \qquad (3.14)$$

III.2.1 Feature Extraction

Feature extraction also plays a very important role in S.I. analysis. In this research a fully automatic S.I. feature extraction algorithm is also proposed. It simply determines the closure and release starting points for medial and final stops, and then computes the duration between these two points.

The first step of this algorithm is to search for the starting point of closure. This point can be found by searching the starting point of articulation until the first empty frame occurs, i.e. the silent period (C or C' in Figure III-6). The second step is to search for the plosive release starting point (R or R' in Figure III-6). It takes the first non-empty frame after the closure starting point, and then computes the S.I. by subtracting the time at the closure starting point from the time at the closure end point. At this moment, the decision on determining whether the stop is a medial stop or final stop has to be made. The algorithm will continue to examine the signal if the voicing period appears after the burst releases. Should it be so, the stop is considered as a medial stop; otherwise, it is a final stop.

with the unknown features and each given class using EQ. 3.15 and EQ. 3.16. Then, the classifier searches for the class which produces the minimum score. Before it assigns the features to that class, it examines them first to see if they meet the conditions of tolerance intervals. The intervals are based on the Chebyshev's Inequality, and are computed by applying equations 3.9-1 and 3.10-1. The following equations are used to compute the discriminant scores for class i.

Medial Stops -

Let $uSI1_i$ and $uSI2_i$ be the average silent intervals of a medial stop in F1 and F2 of class i, and DSIi be the distance between the mean of a stored pattern i and the input features X1 and X2.

$$DSI_i = \sqrt{((X1 - uSI1_i)^2 + (X2 - uSI2_i)^2)} \text{ --------- } (3.15)$$

where i = 1 ..6.

Final Stops -

Let $uSI1'_i$ and $uSI2'_i$ be the average silent intervals of the final stop in F1 and F2 of class i, and DSI'_i be the distance between the mean of a stored pattern i and the input features of X1' and X2'.

$$DSI'_i = \sqrt{((X1' - uSI1'_i)^2 + (X2' - uSI2'_i)^2)} \text{ -- } (3.16)$$

where i = 1 .. 6.

Since the data is in sentence form, a constraint has to be imposed to distinguish S.I. from the durations between words. In this algorithm, a threshold is imposed. If the duration is longer than a given threshold, the detected interval is considered as a word boundary; otherwise, it is accepted as an S.I.

The experiment shows that the algorithm is working well. However, if the word boundary is too narrow, it will be misinterpreted as an S.I. Furthermore, if the silent interval is not well defined, it cannot be detected.

III.2.2 S.I. Classifier

The procedure of this classifier is similar to the F.T. classifier. It also consists of two phases, training and classification. In the training phase, the input data is separated into two halves, one for learning and the other for testing purposes. For the learning data, only the data with stop consonants, is used to form six main groups for medial stops and similarly for final stops. Each of them represents one class of stop consonants. Once all groups have been identified, the next step is to compute the means and standard deviations as well as the tolerance intervals for each group.

The classification phase is the same as F.T.'s, it is achieved through the use of a series of discriminant functions. The discriminant scores are initially computed

III.3 VOICE-ONSET-TIME

V.O.T. is the third method used in this research. It is regarded as the primary cue to distinguish voiced and unvoiced stop consonants. Most researchers like Blumstein and Stevens[25], Eimas and Corbit[26], Lisker and Abramson[27], Lisker, Liberman and Erickson[28], Santerre and Suen[16], Stevens and Klatt[29], Winitz, LaRiviere and Herriman[30], and Wolf[18] conclude that the V.O.T. is usually shorter in voiced stop consonants than in unvoiced stop consonants. V.O.T. is defined as the time interval between the burst that marks the release of the stop closure and the onset of quasi-periodicity which reflects laryngeal vibration (Lisker and Abramson[27]).

This method measures the burst period, i.e. the time difference in the first two formants between: 1) the starting point of plosive release (R) and the starting point of the steady state (S) for the preceding stops, 2) the starting point of plosive release (R') and the ending point of the burst (E) for the final stops. Figure III-7 shows the measurements of V.O.T. The computations of V.O.T. of the preceding and final stops are as follows, where VOT1 & VOT2 and VOT1' & VOT2' represent the V.O.T.'s of the preceding and final stops respectively.

Preceding Stops -

$$VOT1 = s1 - r1 \quad (3.17)$$

$$VOT2 = s2 - r2 \quad (3.18)$$

The steps involved are similar to the F.T. algorithm for the detection of the preceding stop. The only difference is that the latter algorithm is looking for the frequency changes but the former is looking for time differences. However, their plosive release and steady state starting points remain the same. Once these points are found, the algorithm will compute the V.O.T. of the preceding stop using equations 3.17 and 3.18.

The next step is to search for the V.O.T. of the final stop. It searches for the end point of the S.I. and the end point of the following burst. Then, it computes the time difference between these two points using equations 3.19 and 3.20. The distinction between the preceding and the final stops is similar to S.I., which can be found that there is a voicing period following the burst in the preceding stop but not in the final stop.

The results of this experiment also show that the algorithm is working quite well except in the following cases:

(a) The same as the first case of F.T. - the steady state cannot be found in the detection of the preceding stop; and

(b) The word boundary is not clear. It may be misinterpreted as the burst.

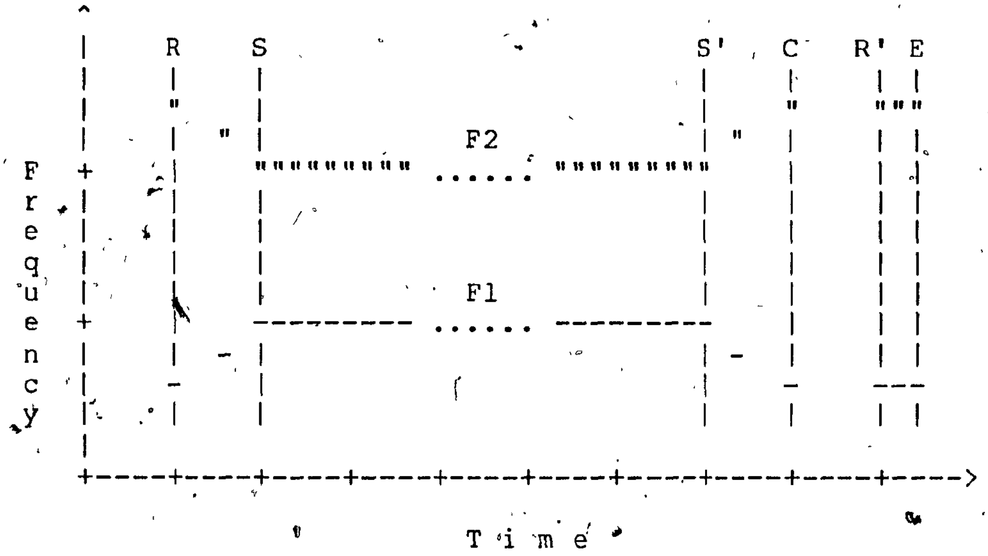


Figure III-7. Measurements of V.O.T.

Final Stops -

$$VOT1' = e1 - r1' \quad (3.19)$$

$$VOT2' = e2 - r2' \quad (3.20)$$

III.3.1 Feature Extraction

A completely automatic V.O.T. feature extraction scheme is proposed in this study. It defines the plosive release starting point and the steady state starting point, and then computes the time difference between these two points for preceding stops. As well the plosive release starting point and the burst end point are defined, and the time difference between these two points is computed for final stops.

III.3.2 V.O.T. Classifier

The V.O.T. classifier consists of training and classification phases. The steps in this classifier are the same as S.I.'s. One half of the stop consonant input data is used to form two six-group (class) reference patterns, one for preceding stops and the other for the classification of final stops. The means and standard deviations of each group are first computed, as are the tolerance intervals using equations 3.9-1 and 3.10-1. The subsequent step is to store the means and tolerance intervals of all classes in the computer memory for classification.

In the classification phase, a set of discriminant functions is used to compute the discriminant scores for every class. The decision on selecting the suitable class is also focused on the one which has the minimum score. Finally, the classifier will check the input feature vectors and find out if they belong to that class. The discriminant score was based on distance measure. Computation formulae for both preceding and final stops are described below:

Preceding stops -

Let $\mathrm{uVOT11}_i$ and $\mathrm{uVOT12}_i$ be the average bursts in F1 and F2 of class i, and DVOT_i be the distance between the mean of a stored pattern i and the input features X1 and X2.

$$\mathrm{DVOT}_i = \sqrt{((X1 - \mathrm{uVOT1}_i)^2 + (X2 - \mathrm{uVOT2}_i)^2)} \text{ ----- } (3.21)$$

where i = 1 .. 6.

Final stops -

Let $uVOT21_i$ and $uVOT22_i$ be the average bursts in F1 and F2 of class i, and $DVOT'_i$ be the distance between the mean of a stored pattern i and the input features X1' and X2'.

$$DVOT'_i = \sqrt{((X1' - uVOT'_i)^2 + (X2' - uVOT2'_i)^2)} \quad - (3.22)$$

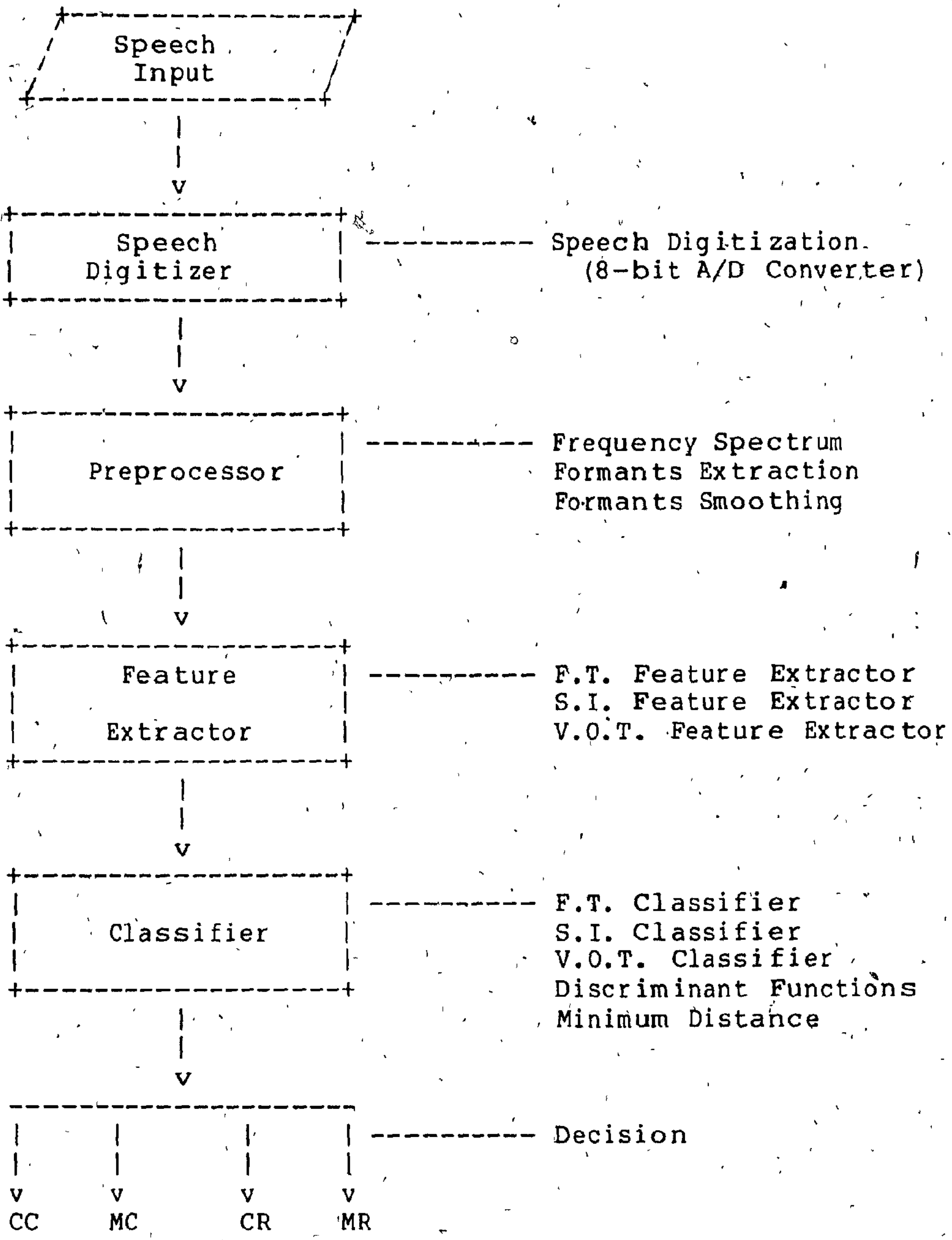
where i = 1 .. 6.

CHAPTER IV

EXPERIMENTAL RESULTS

IV.1 DATA USED IN THE EXPERIMENTS

In order to establish the performance of the proposed CSR system (Detailed structure of which is outlined in Figure IV-1), two different sets of data (half of them were from training set) were prepared for testing the system. They were selected from the speakers with highest recognition scores. The data were chosen equally in number from three male and three female speakers. One set contains seventy-two sentences and the other one contains forty-eight sentences (Originally, there were 72 sentences in this set. Due to a transmission problem, bad data were obtained from two speakers. Therefore, the data of those two speakers are discarded). As mentioned earlier, the difference between these two sets of data is: in the first set, the sentences were uttered by the speakers in their usual way; in the second set, the sentences were uttered by the same speakers but they were requested to emphasize the stop consonants. The respective total number of preceding stop consonants and final stop consonants are 354 (including 138 medial stop consonants) and 228 in unstressed data, and 236 (including 92 medial stop consonants) and 152 in stressed data. The number of stops in each category is shown in detail in APPENDIX III. Each set of data was tested individually by the entire system. Different threshholds were applied. The



Symbols:
CC - Correctly Classified;
MC - Misclassified;
CR - Correctly Rejected;
MR - Misrejected.

Figure IV-1. Detailed Structure of the Proposed CSR System

outcomes of the experiments consist of four types (Figure IV-1). The first type called correct classification (CC), that is the phonemes are correctly classified in the corresponding classes. The second type is called misclassification (MC), i.e. the phonemes are classified in the wrong classes. The third type called correct rejection meaning that the phonemes which do not belong to any classes of stop consonants are rejected by the system. The fourth one called misrejection is that the system incorrectly rejects the phonemes which are actually stop consonants. The following equations are used to compute the rate for each type of result:

$$CC = \frac{\text{Number of correctly classified extracted phonemes}}{\text{Total number of extracted phonemes}}$$

$$MC = \frac{\text{Number of incorrectly classified extracted phonemes}}{\text{Total number of extracted phonemes}}$$

$$= 1 - CC.$$

$$CR = \frac{\text{Number of correctly rejected non-stop consonants}}{\text{Number of extracted phonemes which are non-stops}}$$

$$MR = \frac{\text{Number of incorrectly rejected stop consonants}}{\text{Number of extracted phonemes which are stops}}$$

The detailed results of each method are described in the following sections.

IV.2 RECOGNITION RESULTS OF INDIVIDUAL SPEAKERS

Based on different features used in the CSR system, six classification experiments have been conducted. The results of each experiment show the performance of each CSR subsystem, or each feature used, for each set of data. Table IV-1a shows the recognition scores in percentage for the feature of formant transitions of unstressed data which were obtained from six speakers. The best score in distinguishing single phoneme of speaker 1 is "g" with 92.86%; speaker 2 is "d" with 87.5%; speaker 3 is "p" with 90%; speaker 4 is "b" with 93.33%; speaker 5 is "g" with 81.82% and speaker 6 is "p" with 87.5%. On the other hand, the best score of stressed data is shown in Table IV-1b. The highest rate of speaker 1 is the phoneme "p" with 90%; speaker 3 is "k" with 88.89%; speaker 5 is "g" with 81.82% and speaker 6 is "k" with 80%. On the average, we could obtain the highest scores of all types of phonemes of 74.15% or 109 out of 147 phonemes which are correctly classified from speaker 1 in unstressed data and of 80.71% or 113 out of 140 phonemes from speaker 5 in stressed data. In addition, the misclasification (MC), the correct rejection (CR) and misrejection (MR) rates are 25.85% or 38/147, 75.71% or 53/70 and 19.48% or 15/77 of speaker 1 in unstressed data, and 19.83% or 27/140, 95.77% or 68/71 and 32.81% or 21/69 of speaker 5 in stressed data respectively. The detailed rates for all speakers are described in APPENDIX IV.

It has been found in these experiments that the transitional cues are useful if the target vowels (preceding or following vowels) are known. Moreover, the reference pattern has to be speaker independent, i.e. we cannot apply the same means to any other speakers.

Results of the second feature - silent interval are presented in Table IV-2a and 2b. Again, the best recognition score of unstressed data for single phoneme of speaker 1 is "p" with 100%; speaker 2 is "d", "p" and "k" all with 100%; speaker 3 is "k" with 83.33%; speaker 4 is "d" with 100%; speaker 5 is "b" and "d" both with 100% and speaker 6 is "b", "d" and "p" all with 100%. In stressed data, speaker 1 is 85.71% on phoneme "p"; speaker 3 is 57.14% also on "p"; speaker 5 is 100% on "d" and speaker 6 is 80% on "p". On the average, the best score is speaker or 6 in both unstressed and stressed data with 58.33% or 42/72 and 58.25% 60/103 respectively. The details can be referred to APPENDIX V.

It has also been found in these experiments that the duration of silent interval of the same phoneme is dependent on the location of the word in the sentence. The duration of the last word of the sentence is on the average much longer than anywhere else in the sentence. On the other hand, the length of silent interval of word in the beginning or middle of a sentence is always dependent on the pronunciation habit of the speaker and the context. Figures IV-2 and IV-3 show that different durations will be obtained if the positions of

Spkr	b	d	g	p	t	k	CR	MR	CC
1	75	55	92.86	80	53.33	81.82	75.71	19.48	74.15
2	73.33	87.50	83.33	81.82	53.85	75	65.52	15.49	70.54
3	73.33	50	66.67	90	60	57.14	75	17.19	70.97
4	93.33	70	72.43	75	70	90	52.70	2.94	65.49
5	80	71.43	81.82	80	58.33	77.78	61.67	20.31	68.55
6	66.67	77.78	84.63	87.50	46.67	83.33	75.47	23.61	73.60

Symbols: CC – Correctly Classified; MR - Misrejected.
CR – Correctly Rejected;

Table IV-1a. Recognition scores (%) for the Feature of F.T. Obtained from Six Speakers (Unstressed Data)

Spkr	b	d	g	p	t	k	CR	MR	CC
1	64.29	60	70	90	53.33	71.43	88.76	22.73	79.35
3	80	45.45	84.62	75	46.15	88.89	85.19	28.77	77.92
5	58.82	70	81.82	66.67	46.15	77.78	95.77	32.81	80.71
6	76.47	75	78.57	66.67	33.33	80	91.03	25.71	80.41

Symbols: CC - Correctly Classified; MR - Misrejected.
CR - Correctly Rejected;

Table IV-1b. Recognition Scores (%) for the Feature of F.T. Obtained from Four Speakers (Stressed Data)

Spkr	b	d	g	p	t	k	CR	MR	CC
1	40	66.67	50	100	41.67	100	43.64	15.91	52.53
2	66.67	100	75	100	62.50	100	29.79	4.17	46.48
3	0	75	66.67	0	62.50	83.33	15.22	3.85	33.33
4	50	100	40	0	28.57	71.43	31.48	11.11	38.27
5	100	100	75	75	0	66.67	50.98	21.74	55.41
6	100	100	62.50	100	11.11	66.67	60.47	17.24	58.33

Symbols: CC - Correctly Classified; MR - Misrejected.
CR - Correctly Rejected;

Table IV-2a. Recognition Scores (%) for the Feature of S.I. Obtained from Six Speakers (Unstressed Data)

Spkr	b	d	g	p	t	k	CR	MR	CC
1	60	57.14	33.33	85.71	43.75	50	37.50	7.69	43.97
3	50	55.56	50	57.14	47.06	50	58.97	12.28	55.56
5	60	100	50	66.67	46.67	62.50	52.54	6.52	56.19
6	66.67	50	62.50	80	58.33	50	56.92	7.89	58.25

Symbols: CC - Correctly Classified; MR - Misrejected.
CR - Correctly Rejected;

Table IV-2b. Recognition Scores (%) for the Feature of S.I. Obtained from Four Speakers (Stressed Data)

DATE - 81/09/23.   TIME - 18.37.51.

BILL HERTHA- "HE WENT BACK INTO THE DEEP FOREST."

```
    0.00  .39  .78 1.17 1.56 1.95 2.34 2.73 3.13 (KHZ)
    0. |----|----|----|----|----|----|----|----|--->
 12.80 |
 25.60 |
 38.40 |  .                         "
 51.20 |  .                        "
 64.00 +  .                        "
 76.80 |  .                        "
 89.60 |  .                       "
102.40 |  .                      "
115.20 |  .                      "
128.00 +  .                   "
140.80 |  .          "
153.60 |   .         "
166.40 |   .         "
179.20 |    .        "
192.00 +     .       "
204.80 |      .        "
217.60 |      .          "
230.40 |      .           "
243.20 |      .             "
256.00 +      .              "
268.80 |     .               "
281.60 |    .              "
294.40 |   .         "
307.20 |  .        "
320.00 +  .        "
332.80 |  .        "
345.60 |
358.40 |     .                "
371.20 |    .                 "
384.00 +
396.80 |
409.60 |
422.40 |
435.20 |
448.00 +
       .     where . represents the first formant;
       .           " represents the second formant.
       .
       V
```

Figure IV-2. The Silent Duration of the Phoneme "t" of the Word "went" Starts From the Time Frame 345.60 to 358.40

the word are different. The example given here is for the phoneme "t". The first "t" that appears in Figure IV-2 is located at the second place of the sentence. However, the second "t" is located at the end of the sentence (Figure IV-3).

Tables IV-3a and 3b present the recognition results of the third feature - voice onset time. Similarly, the best recognition score of unstressed data for single phoneme of different speakers are as following: speaker 1 and 2 are phoneme "b" with 75% and 62.5%; speaker 3 and 4 are "d" with 63.64% and 62.5%; speaker 5 is "b" and "k" both with 62.5% and speaker 6 is "k" with 55.56%. In stressed data, the best rate for single phoneme of all speakers is phoneme "b" except for speaker 5 is phoneme "d" with 75.67%. Overall, the best scores are 50% from three speakers 1 (65/130), 3 (60/120) and 4 (63/126) in unstressed data, and 56.79% from speaker 5 (74/131) in stressed data. The details can be seen in APPENDIX VI.

The results of these experiments indicate that the existing problems are similar to the silent interval's, i.e. the burst period as well as the closure period, is completely dependent on the context (the environment of the target word) and the usual way the speakers pronounce them.

Tables IV-4a and 4b summarize the scores of correct rejection, misrejection and correct classification of the

```
        DATE - 81/09/23.     TIME - 18.37.51.

        BILL HERTHA- "HE WENT BACK INTO THE DEEP FOREST.".

     0.00  .39  .78 1.17 1.56 1.95 2.34 2.73 3.13 (KHZ)
     0. |----|----|----|----|----|----|----|----|--->
        .
1638.40 |
1651.20 |        .   "
1664.00 +        .   "
1676.80 |       .    "
1689.60 |       .    "
1702.40 |      .     "
1715.20 |     .      "
1728.00 +     .       "
1740.80 |     .       "
1753.60 |     .       "
1766.40 |     .        "
1779.20 |     .        "
1792.00 +    .           "
1804.80 |    .           "
1817.60 |    .            "
1830.40 |    .          "
1843.20 |    .      "
1856.00 +    .     "
1868.80 |    .     "
1881.60 |
1894.40 |
1907.20 |          .
1920.00 +               "
1932.80 |     .            "
1945.60 |     .                 "
1958.40 |     .                  "
1971.20 |     .                      "
1984.00 +   .                        "
1996.80 |   .                        "
2009.60 |   .                     "
2022.40 |   .              "
2035.20 |   .              "
2048.00 +
2060.80 |
2073.60 |
2086.40 |
2099.20 |
2112.00 +
2124.80 |    .                          "
2137.60 |                               "
2150.40 |                               "
2163.20 |
        V
```

Figure IV-3. The Silent Duration of the Phoneme "t" of the Word "forest" Starts From the Time Frame 2048.00 to 2124.80

Spkr	F.T.			S.I.			V.O.T.		
	CR	MR	CC	CR	MR	CC	CR	MR	CC
1	75.71	19.48	74.15	43.64	15.91	52.53	45.61	5.48	50
2	65.52	15.49	70.54	29.79	4.17	46.48	41.86	4.48	46.67
3	75	17.19	70.97	15.22	3.85	33.33	42.59	7.58	50
4	52.70	2.94	65.49	31.48	11.11	38.27	46.03	4.76	50
5	61.67	20.31	68.55	50.98	21.74	55.41	41.18	7.94	50
6	75.47	23.61	73.60	60.47	17.24	58.33	41.86	1.64	46.15

Symbols: CC - Correctly Classified; MR - Misrejected.
CR - Correctly Rejected;

Table IV-4a. Summary of Recognition Scores (%) for All Features (Unstressed)

Spkr	F.T.			S.I.			V.O.T.		
	CR	MR	CC	CR	MR	CC	CR	MR	CC
1	88.76	22.73	79.35	37.50	7.69	43.97	46.97	1.14	52.55
3	85.19	28.77	77.92	58.97	12.28	55.56	46.15	1.33	52.14
5	95.77	32.81	80.71	52.54	6.52	56.19	48.44	2.99	56.49
6	91.03	25.71	80.41	56.92	7.89	58.25	45	1.30	54.74

Symbols: CC - Correctly Classified; MR - Misrejected.
CR - Correctly Rejected;

Table IV-4b. Summary of Recognition Scores (%) for All Features (Stressed)

Spkr	b	d	g	p	t	k	CR	MR	CC
1	75	50	45.45	50	47.06	45.45	45.61	5.48	50
2	62.50	42.86	50	44.44	45.45	44.44	41.86	4.48	46.67
3	61.54	63.64	40	55.56	60	50	42.59	7.58	50
4	56.25	62.50	45.45	55.56	54.55	50	46.03	4.76	50
5	62.50	55.56	45.45	55.56	45.45	62.50	41.18	7.94	49.12
6	50	50	50	42.86	46.15	55.56	41.86	1.64	46.15

Symbols: CC - Correctly Classified; MR - Misrejected.
CR - Correctly Rejected;

Table IV-3a. Recognition Scores (%) for the Feature V.O.T. Obtained from Six Speakers (Unstressed Data)

Spkr	b	d	g	p	t	k	CR	MR	CC
1	69.23	58.33	55.56	58.33	43.75	66.67	46.97	1.41	52.55
3	64.29	57.14	57.14	58.33	52.63	55.56	46.15	1.33	52.14
5	69.23	75.67	54.55	54.55	53.33	62.50	48.44	2.99	56.49
6	93.75	62.5	61.54	50	52.63	45.45	45	1.30	54.74

Symbols: CC - Correctly Classified; MR - Misrejected.
CR - Correctly Rejected;

Table IV-3b. Recognition Scores (%) for the Feature of V.O.T. Obtained from Four Speakers (Stressed Data)

three distinctive features for all speakers including both unstressed and stressed data.

IV.3 RECOGNITION RESULTS OF ALL SPEAKERS

The global results of all speakers can be seen in Tables IV-5 to IV-7. Table IV-5a shows the confusion matrix of the unstressed data. Altogether 792 Formant Transition features were extracted from six speakers of which 559 were correctly recognized. Each of them uttered 12 sentences. The most confused phoneme is "t". Only 43 out of 76 (i.e. the total number of features on the row "t"), features can be identified. The recognition rate of each phoneme is listed on the rightmost column - R.Rate. In the other set of data, the overall performance is superior to the first set. Results are based on four selected speakers each uttered 12 sentences (Table IV-5b). There were 597 F.T. features extracted and tested by the F.T. subsystem which scored 79.56%. Similarly, the most difficult phoneme to detect is "t" with only 45.28% recognition rate.

The results of the second distinctive feature, silent interval, are shown in Tables IV-6a and 6b. There were 469 features extracted from the unstressed data, and 459 features extracted from the stressed data in these experiments. It may be noted that the number of features in the former set of data is much smaller than the latter one if we divide the number of

Speaker		Classified						1	2
All		b	d	g	p	t	k	other	R.Rate
Actual	b	70	0	7	2	0	0	12	76.92
	d	4	37	2	1	0	0	11	67.27
	g	0	1	59	2	0	2	9	80.82
	p	2	0	2	50	1	0	6	81.97
	t	2	1	5	1	43	1	23	56.58
	k	1	1	2	2	0	48	7	78.69
	other	37	6	27	14	15	24	252	67.2
Average								559/792	= 70.58

Table IV-5a. Global Recognition Results for the Feature of F.T. Obtained from Six Speakers (Unstressed)

Speaker		Classified						1	2
All		b	d	g	p	t	k	other	R.Rate
Actual	b	44	2	1	1	0	0	15	69.84
	d	4	24	0	0	0	0	11	61.54
	g	0	0	38	1	0	1	8	79.17
	p	2	0	1	30	0	0	8	75
	t	1	0	0	0	24	0	28	45.28
	k	1	0	1	0	0	28	5	80
	other	10	12	5	1	2	2	287	89.97
Average								475/597	= 79.56

Table IV-5b. Global Recognition Results for the Feature of F.T. Obtained from Four Speakers (Stressed)

Note: 1 Non-stop-consonants; 2 Recognition rate.

	Speaker	Classified							
	All	b	d	g	p	t	k	other[1]	R.Rate[2]
Actual	b	10	1	1	0	2	2	2	55.56
	d	0	17	0	0	1	1	1	85
	g	3	1	21	2	1	3	4	60
	p	0	0	0	12	1	0	1	85.71
	t	4	5	2	1	18	5	13	37.5
	k	1	2	2	0	1	31	1	81.58
	other	10	47	28	20	37	40	114	38.51
	Average								223/469 = 47.55

Table IV-6a. Global Recognition Results for the Feature of S.I. Obtained from Six Speakers (Unstressed)

	Speaker	Classified							
	All	b	d	g	p	t	k	other[1]	R.Rate[2]
Actual	b	11	1	0	3	2	1	1	57.89
	d	1	16	2	0	1	2	2	66.67
	g	3	3	15	2	7	0	1	48.39
	p	0	0	1	18	1	0	5	72
	t	6	6	5	6	29	4	4	48.33
	k	4	1	3	3	1	18	4	52.94
	other	10	39	13	33	14	19	138	51.88
	Average								245/459 = 53.38

Table IV-6b. Global Recognition Results for the Feature of S.I. Obtained from Four Speakers (Stressed)

Note: 1 Non-stop-consonants; 2 Recognition rate.

	Speaker	Classified							
	All	b	d	g	p	t	k	other[1]	R.Rate[2]
Actual	b	56	11	7	2	3	5	7	61.54
	d	11	29	7	0	2	3	1	54.72
	g	10	7	29	5	3	4	4	46.77
	p	8	4	4	26	0	7	2	50.98
	t	9	9	5	4	39	8	4	50
	k	6	5	3	5	4	27	3	50.94
	other	61	42	22	15	15	21	135	43.41
	Average								341/699 = 48.78

Table IV-7a. Global Recognition Results for the Feature of V.O.T. Obtained from Six Speakers (Unstressed)

	Speaker	Classified							
	All	b	d	g	p	t	k	other[1]	R.Rate[2]
Actual	b	42	2	2	5	1	3	1	75
	d	4	22	1	2	5	1	1	61.11
	g	8	2	27	2	3	5	0	57.45
	p	7	4	2	25	3	4	0	55.56
	t	11	2	2	5	35	12	2	50.72
	k	3	3	2	5	2	21	1	56.76
	other	56	14	10	27	17	12	119	52.89
	Average								291/515 = 56.50

Table IV-7b. Global Recognition Results for the Feature of V.O.T. Obtained from Four Speakers (Stressed)

Note: 1 Non-stop-consonants; 2 Recognition rate.

features by the number of speakers. The experiments show that some final stops do not exist in the data because the speakers did not articulate them clearly. It can be concluded that the speakers generally ignore the pronunciation of the final stops in their normal conversation. On the average, the recognition scores of the unstressed data and stressed are in the order of 47.55% and 53.38%. The reasons why the results are worse compared with the previous two sets of results are partly due to the difficulty in distinguishing between silent durations and word boundaries, and partly because the silent interval is always affected by its position in the word, and those words preceding or following the target word.

The results of the third method - voice onset time, are presented in Table IV-7a for unstressed data and Table IV-7b for stressed data respectively. The total number of features extracted via this method is 699 for unstressed and 515 for stressed data. Only 341 of them are identified correctly from unstressed data and so are 291 from stressed data. The average scores are 48.78% for the unstressed data and 56.50% for the stressed data. The results are not so good because the variations of the data are too big. The burst period for most of the phonemes varied from the minimum of 12.8 msec to high of 128 msec range. This incurs difficulties on creating the reference pattern or in training the classifier.

Tables IV-8a and 8b show the overall performance of each recognition method. In unstressed data, the best recognition

Method	F.T.	S.I.	V.O.T.
CR	252/375=67.20%	114/296=38.51%	135/311=43.41%
MR	68/417=16.31%	[illegible]2/173=12.72%	21/388= 5.41%
CC	559/792=70.58%	223/469=47.55%	341/699=48.78%
MC	233/792=29.42%	246/469=52.45%	358/699=51.22%

Symbols: CC - Correctly Classified; MC - Misclassified;
CR - Correctly Rejected; MR - Misrejected.

Table IV-8a. Comparative Recognition Results of Three Distinctive Features (Unstressed)

Method	F.T.	S.I.	V.O.T.
CR	287/319=89.97%	138/266=51.88%	119/255=46.67%
MR	75/278=26.98%	17/193= 8.81%	5/260= 1.92%
CC	475/597=79.56%	245/459=53.38%	291/515=56.50%
MC	122/597=20.44%	214/459=46.62%	224/515=43.50%

Symbols: CC - Correctly Classified; MC - Misclassified;
CR - Correctly Rejected; MR - Misrejected.

Table IV-8b. Comparative Recognition Results of Three Distinctive Features (Stressed)

CHAPTER V

CONCLUSIONS AND SUGGESTIONS

V.1 CSR SYSTEM REVIEW

A speaker independent speech recognition system for the automatic recognition of stop consonants in continuous speech has been developed. The entire system consists of four stages. The first stage called digitization which contains the components of one 8-bit A/D converter and one micro computer system. Its function is to convert the speech signal into digital form for computer processing. The second stage is called preprocessing. It takes place in two phases. Firstly, the time domain digital signals are transformed into the frequency domain speech spectra using the discrete fast Fourier transform method. Then, the next phase is to extract the characteristics of the movements of the vocal tract called formants through peak-picking method. Smoothing by means of interpolation is applied simultaneously in order to obtain smooth formants. Feature extraction is the next stage of the system. Three different features called formant transitions, silent interval and voice-onset-time are extracted individually via three feature extractors. Once the features have been extracted, they are sent to the final stage - classification for identification. Each type of feature is examined by its corresponding classifier for identification. The techniques used in the classifiers are mainly statistical.

method is Formant Transitions which obtain 70.58% compare with other two methods which obtain below 50%. In stressed data, the best method is also Formant Transitions which obtain 79.56% compare with 53.38% of Silent Interval method and 56.50% of Voice-Onset-Time method.

A set of discriminant functions is first computed. Based on the minimum distance classification technique, the input feature is classified into the class with a minimum distance. A threshold called tolerance interval is imposed in order to determine if the classified phoneme belongs to that class.

V.2 CONCLUSIONS

The performance of each method used in the present system has already been discussed. The results show that no single feature alone can perfectly account for the distinction of each voiced and unvoiced stop consonants. The formant transition cue produced the highest recognition score only with a priori knowledge of the target vowels. The other two cues are mainly time considerations. As noted in the previous chapter, low recogniton rates are due to considerable variations of burst or silent period when the position of the word is different. On the other hand, some speakers ignored the pronunciation of some phonemes especially the final stops. As a result, some final stops cannot be detected. However, this can be compensated by using the transitional cues.

Conclusion can also be drawn that timing would not be an effective cue to measure the phonemes in continuous speech unless certain constraints can be imposed, such as the speed and intonation of the speaker who speaks the same word which appears in different positions should be kept the same, and phonemes, we can combine these features for classifications to obtain better recognition results. For example, if a final stop follows a nasal, it cannot be easily detected by the Formant Transition extractor because the transition rate is too high. As a result, not much information can be extracted. However, it can be detected by the Silent Interval or Voice-Onset-Time extractor. Since the feature extractor cannot extract some unusual features such as the cases listed in each feature extraction section in Chapter III, the suggested method may provide mutual benefits.

the speakers should pay attention to the pronunciation of the final stops.

V.3 SUGGESTIONS FOR FURTHER STUDY

In order to obtain higher performance of the proposed system, some details for further development in ths system are suggested and outlined as follows:

1). In the system, only the stop consonants are considered. It can be extended to recognize other phonemes such as vowels and other consonants.

2). There is a difficulty encountered in splitting a long speech during digitization, because the silent gap is too small to divide it manually. This can be solved by modifying the digitization system to detect the silent gap by itself.

3). Almost 50% of the complete processing time is used up by the DFFT algorithm. The processing time can be reduced if the algorithm can ignore processing the gaps between words. This can be achieved if the word boundary can be detected.

4). The problems of distinguishing the silent interval and word boundaries during feature extraction can be solved by looking at the energy concentrations. Usually there is less energy at the end of the word than at the middle of the word.

5). Instead of using only single features to detect the

REFERENCES

[1]. Potter, R.K., Kopp, G.A. and Kopp, H.G., "Visible Speech", 1947, (Reprinted New York, 1966).

[2]. Drey-Grof, J., "Sonograph and Sound Mechanics", J. Acoust. Soc. Am., Vol.24, pp.731-739, 1950.

[3]. Davis, K.H., Biddulph, K.H., Balashek, S., "Automatic Recognition of Spoken Digits", J. Acoust. Am., Vol.24, pp.637-642, 1952.

[4]. Fry, D.B. and Denes, P.B., "Mechanical Speech Recognition", Communication Theory, London, Butterworth, pp.426-432, 1953.

[5]. Rabiner, L.R., Schafer, R.W., "Digital Processing of Speech Signals", Prentice-Hall Inc. P.43, 1978.

[6]. Markel, J.D., "Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation", IEEE Trans. Audio Electro Acoust., Vol.AU-20, pp.129-137, June, 1977.

[7]. Demichelis, P., De Mori R., Laface, P. and O'Kane, M., "Computer Recognition of Stop Consonants", IEEE International Conference on Acoust., Speech and Signal Processing, pp.85-88, 1979.

[8]. Itakura, F., "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Trans. Acoust., Speech and Signal processing, Vol.ASSP-23, pp.67-72, Feb.1975.

[9]. Molho, L., "Automatic Acoustic-Phonetic Analysis of Fricatives and Plosives", IEE Acoust. Speech Signal Process. Rec., pp.182-185, Apr.1976.

[10]. Searle, C.L., Jacobson, J.Z. and Rayment, S.G., "Stop Consonant Discrimination Based on Human Audition", J. Acoust. Soc. Am., Vol.65, No.3, pp.799-809, Mar.1979.

[11]. Weinstein, C.J., McCandless, S.S., Mondshein, L.F. and Zue V.W., "A System for Acoustic-Phonetic Analysis of Continuous Speech", IEEE Trans. Acoust., Speech and Signal Process., Vol.ASSP-23, pp.314-327, Feb.1975.

[12]. Cole, R.A. and Scott, B., "Toward a Theory of Speech Perception", Psychologic Reveiw, Vol.81, No.4, pp.348-374, 1974.

[13]. Datta, A.K., Ganguli, N.R. and Ray, S., "Recognition of Unaspirated Plosives - A Statistical Approach", IEEE

[26]. Eimas, P.D. and Corbit, J.D., "Selective Adaptation of Linguistic Feature Detectors", Cognitive Psychology 4, pp.99-109, 1973.

[27]. Lisker, L. and Abramson, A.S., "Some Effects of Context on Voice Onset Time in English Stops", Language and Speech 10, pp.1-28, 1967.

[28]. Lisker, L., Liberman, A.M. and Erickson, D.M., "On Pushing the Voice-Onset-Time (VOT) Boundary About", Langyage and Speech 20, pp.209-216, 1977.

[29]. Stevens, K.N. and Klatt, D.H., "Role of Formant Transitions in Voiced-Voiceless Distinction for Stops", J. Acoust. Soc. of Am., Vol.55, No.3, pp.653-659, Mar.1974.

[30]. Winitz, H., LaRiviere, C. and Herriman, E., "Variations in VOT for English Initial Stops", J. of Phonetics, Vol.3, pp.41-52, 1975.

[31]. Suen, C.Y., "n-Gram Statistics for Natural Language Understanding and Text Processing", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.PAM-1, No.2, pp.164-172, Apr.1979.

[32]. Flanagan, J.L., "Automatic Extraction of Formant Frequencies from Continuous Speech", J. Acoust. Soc. of Am., Vol.28 pp.110-118, Jan.1956.

[33]. Bergland, G.D., "A Guided Tour of the Fast Fourier Transform", IEEE Spectrum, pp.41-52, July,1969.

[34]. Markel, J.D. and Gray, G.H. Jr., "Linear Prediction of Speech", New York, Springer-Verlag, 1976.

[35]. Oppenheim, A.V., "Speech Spectrograms Using the Fast Fourier Transform", IEEE Spectrum, pp.57-62, Aug.1970.

[36]. Schafer, R.W. and Rabiner, L.R., "System for Automatic Formant Analysis of Voiced Speech", J. Acoust. Am., Vol.47, pp.634-648, Feb.1970.

[37]. McCandless, S.S., "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra", IEEE Trans. on Acoust., Speech & Signal Process. Vol.ASSP-22, pp.135-141, Apr.1974.

[38]. Duda, R.O. and Hart, P.E., "Pattern Classifier and Scene Analysis", Wiley-Interscience, 1973.

[39]. Gupta, S.C. and Kapoor, V.K., "Fundamentals of Mathematical Statistics", Sultan Chand & Sons, 7th ed., p.307, Jan.1980.

Trans. on Acoust., Speech and Signal Process., Vol.ASSP-28, No.1, pp.85-91, Feb.1980.

[14]. Menon, K.M.N., Rao, P.V.S. and Thosar, R.B., "Formant Transitions and Stop Consonant Perception in Syllables", Language and Speech 17, pp.27-46, 1974.

[15]. Pal, S.K. and Majumder, D.D., "Effect of Fuzzification on the Plosive Cognition System", Int. J. Systems Sci., Vol.9, No.8, pp.873-886, 1978.

[16]. Santerre, L. and Suen C.Y., "Why Look for a Single Feature to Distinguish Stop Cognates ?", Journal of Phonetics 9, pp.163-174, 1981.

[17]. Sharf, D.J. and Hemeyer, T., "Identification of Place of Consonant Articulation from Vowel Formant Transition", J. Ascoust. Soc. Am. Vol.51, No.2 (Part 2) pp.652-658, 1972.

[18]. Wolf, C.G., "Voicing Cues in English Final Stops", J. of Phonetics, Vol.6, 19, pp.299-309, 1978.

[19]. Liberman, A., Harris, K.S., Eimas, P., Lisker, L. and Bastian, J., "An effect of learning on speech perception: the discrimination of durations of silence with and without phonemic significance", Language and Speech, 4, p.175, 1961.

[20]. Lisker, L., "Closure Duration and the Intervocalic Voiced-Unvoiceless Distinction in English", Language 33, pp.42-49, 1957.

[21]. Port, R.F., "The Influence of Tempo on Stop Closure Duration as a Cue for Voicing and Place", J. of Phonetics, Vol.7, pp.45-56, 1979.

[22]. Slis, I.H. and Cohen, A., "On the Complex Regulating the Voiced-voiceless Distinction I", Language and Speech 12, p.80, I969.

[23]. Slis, I.H. and Cohen, A., "On the Complex Regulating the Voiced-voiceless Distinction II", Language and Speech I2, p.137, 1969.

[24]. Suen, C.Y. and Beddoes, M.P., "The Silent Interval of Stop Consonants", Language and Speech 17, pp.126-134, Apr-June, 1974.

[25]. Blumstein, S.E. and Stevens, K.N., "Perceptual Invariance and Onset Spectra for Stop Consonants in Different Vowel Environments", J. Acoust. Soc. Am., Vol.67, pp.648-662, Feb.1980.

APPENDIX I

* 1. BOB PAINTED THE BODY OF HIS CAR GREEN.

* 2. HE WENT BACK INTO THE DEEP FOREST.

* 3. THE GAP BETWEEN THEORY AND PRACTICE IS BIG.

4. SHE NEVER EXPLAINED WHY SHE CRIED.

* 5. HE PUT THE SLEEPING DOG IN THE LAB.

6. THE SPARE-RIB IS BETTER THAN THE STEAK.

* 7. NOBODY KNEW THE PRINCE TOOK A CAB AND LEFT.

8. STRIKE OF INDEPENDENT TRUCKERS WILL END SOON.

9. FRESH CAPITAL IS REQUIRED TO MODERNIZE OLD PLANTS.

* 10. JACK GOT A GOLD MEDAL AGAIN FOR HIS BOOK.

11. THE TROUBLE-MAKER WAS TAKEN AWAY.

12. DRIVERS IN NEW YORK WAIT FOR HOURS TO BUY GAS.

* 13. DOUG IS STILL LAGGING BEHIND UP THE HILL.

14. THE REBELLIOUS MOB DISPERSED AFTER THE POLICE CAME.

* 15. JOHN TOOK A DEEP BREATH BEFORE HE WALKED ON THE ROPE.

* 16. HAIG WILL BE A PARTICIPANT IN SALT TWO DEBATE.

17. THE ANGRY SINGER HIT THE MIKE WITH A GUITAR.

* 18. A NEW GOVERNMENT WAS SET UP TWO WEEKS AGO.

* 19. THE WORKER DUG A HOLE TO LET THE WATER RUN.

* 20. BOBBY BROKE HIS ANKLE WHILE PLAYING IN THE GARDEN.

* Sentences tested by the system.

APPENDIX II

COMBINATIONS OF CV AND VC TABLE :

	b	d	g	p	t	k
ae	1 2	2	1	1 2		1 2
eI			2	1		
e	1					2
a	1 2		1	1	2	1
ɛ	1				2	
i	1	1	1	2	1	2
I	1	1	2	1	1	
o	1	2	1	2		2
ɔ		1	2		2	
u	1			1	1 2	2
ʌ		1	1 2	2		
ə	1	1 2	1	1	1 2	1
l		1				1
aI		2				

Symbols -

1 - Stop consonant followed by a vowel (CV);

2 - Vowel followed by a stop consonant (VC).

APPENDIX III

NUMBER OF STOP CONSONANTS IN EACH CATEGORY* :

Type	b	d	g	p	t	k	Total
CV	15	10	8	8	13	5	59
VC	3	6	5	6	11	7	38
Total	18	16	13	14	24	12	97

Symbols -

CV - Preceding stop consonants;

VC - Final stop consonants.

* The numbers in the table are based on one speaker. Therefore, in order to obtain the total number of stops in each category of each set of data, each number has to be multiplied by 6 for unstressed data and multiplied by 4 for stressed data.

APPENDIX IV

Spkr	CR	MR	CC	MC
1	53/70=75.71	15/77=19.48	109/147=74.15	38/147=25.85
2	38/58=65.52	11/71=15.49	91/129=70.54	38/129=29.46
3	45/60=75.00	11/64=17.19	88/124=70.97	36/124=29.03
4	39/74=52.70	2/68= 2.94	93/142=65.49	49/142=34.51
5	37/60=61.67	13/64=20.31	85/124=68.55	39/124=31.45
6	40/53=75.47	17/72=23.61	92/125=73.60	33/125=26.40

Symbols: CC - Correctly Classified; MC - Misclassified;
CR - Correctly Rejected; MR - Misrejected.

(a). Recognition Scores (%) for the Feature of F.T. Obtained from Six Speakers (Unstressed Data)

Spkr	CR	MR	CC	MC
1	79/89=88.76	15/66=22.73	123/155=79.35	32/155=20.65
3	69/81=85.19	51/73=28.77	120/154=77.92	34/154=22.08
5	68/71=95.77	21/69=32.81	113/140=80.71	27/140=19.29
6	71/78=91.03	18/70=25.71	119/148=80.41	29/148=19.59

Symbols: CC - Correctly Classified; MC - Misclassified;
CR - Correctly Rejected; MR - Misrejected.

(b). Recognition Scores (%) for the Feature of F.T. Obtained from Four Speakers (Stressed Data)

APPENDIX V

Spkr	CR	MR	CC	MC
1	24/55=43.64	7/44=15.91	52/99=52.53	47/99=47.47
2	14/47=29.79	1/24= 4.17	33/71=46.48	38/71=53.52
3	7/46=15.22	1/26= 3.85	24/72=33.33	48/72=66.67
4	17/54=31.48	3/27=11.11	31/81=38.27	50/81=61.73
5	26/51=50.98	5/23=21.74	41/74=55.41	33/74=44.59
6	26/43=60.47	5/29=17.24	42/72=58.33	30/72=41.67

Symbols: CC - Correctly Classified; MC - Misclassified;
CR - Correctly Rejected; MR - Misrejected.

(a). Recognition Scores (%) for the Feature of S.I. Obtained from Six Speakers (Unstressed Data)

Spkr	CR	MR	CC	MC
1	24/64=37.50	4/52= 7.69	51/116=43.97	65/116=56.03
3	46/78=58.97	7/57=12.28	75/135=55.56	60/135=44.44
5	31/59=52.54	3/46= 6.52	59/105=56.19	46/105=43.81
6	37/65=56.92	3/38= 7.89	60/103=58.25	43/103=41.75

Symbols: CC - Correctly Classified; MC - Misclassified;
CR - Correctly Rejected; MR - Misrejected.

(b). Recognition Scores (%) for the Feature of S.I. Obtained from Four Speakers (Stressed Data)

APPENDIX VI

Spkr	CR	MR	CC	MC
1	26/57=45.61	4/73= 5.48	65/130=50.00	65/130=50.00
2	18/43=41.86	3/62= 4.48	49/105=46.67	56/105=53.33
3	23/54=42.59	5/66= 7.58	60/120=50.00	60/120=50.00
4	29/63=46.03	3/63= 4.76	63/126=50.00	63/126=50.00
5	21/51=41.18	5/63= 7.94	56/114=49.12	58/114=50.88
6	18/43=41.86	1/61= 1.64	48/104=46.15	56/104=53.85

Symbols: CC - Correctly Classified; MC - Misclassified;
CR - Correctly Rejected; MR - Misrejected.

(a). Recognition Scores (%) for the Feature of V.O.T. Obtained from Six Speakers (Unstressed Data)

Spkr	CR	MR	CC	MC
1	31/66=46.97	1/71= 1.41	72/137=52.55	65/137=47.45
3	30/65=46.15	1/75= 1.33	73/140=52.14	67/140=47.86
5	31/64=48.44	2/67= 2.99	74/131=56.49	57/131=43.51
6	27/60=45.00	1/77= 1.30	75/137=54.74	62/137=45.26

Symbols: CC - Correctly Classified; MC - Misclassified;
CR - Correctly Rejected; MR - Misrejected.

(b). Recognition Scores (%) for the Feature of V.O.T. Obtained from Four Speakers (Stressed Data)