INFORMATION TO USERS

This reproduction was made from a copy of a manuscript sent to us for publication and microfilming. While the most advanced technology has been used to photograph and reproduce this manuscript, the quality of the reproduction is heavily dependent upon the quality of the material submitted. Pages in any manuscript may have indistinct print. In all cases the best available copy has been filmed.

The following explanation of techniques is provided to help clarify notations which may appear on this reproduction.

1. Manuscripts may not always be complete. When it is not possible to obtain missing pages, a note appears to indicate this.

2. When copyrighted materials are removed from the manuscript, a note appears to indicate this.

3. Oversize materials (maps, drawings, and charts) are photographed by sectioning the original, beginning at the upper left hand corner and continuing from left to right in equal sections with small overlaps. Each oversize page is also filmed as one exposure and is available, for an additional charge, as a standard 35mm slide or in black and white paper format.*

4. Most photographs reproduce acceptably on positive microfilm or microfiche but lack clarity on xerographic copies made from the microfilm. For an additional charge, all photographs are available in black and white standard 35mm slide format.*

*For more information about black and white slides or enlarged paper reproductions, please contact the Dissertations Customer Services Department.

UMI Dissertation Information Service

University Microfilms International
A Bell & Howell Information Company
300 N. Zeeb Road, Ann Arbor, Michigan 48106
AUTOMATIC IDENTIFICATION AND RECOGNITION OF DEAF SPEECH

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in the Graduate School of the Ohio State University

By

Kadry A. Abdelhamied, B.S., M.S.,
The Ohio State University

1986

Dissertation Committee:
Dr. M.B. Waldron
Dr. R.A. Fox
Dr. P.S. Niswander

Approved by

Adviser
Department of Biomedical Engineering
To my mother, with love and gratitude.
ACKNOWLEDGEMENTS

I would like to acknowledge my indebtedness to my adviser, Dr. M.B. Waldron, for her guidance and constant support throughout the formulation and presentation of this study. I would like to thank the other members of my advisory committee, Dr. R.A. Fox and Dr. P.S. Niswander for their valuable suggestions and comments. My thanks also go to Prof. H. Weed for his useful comments.

Many thanks are due to Dr. W. Cowart, Dr. R. Fox, Keith Johnson and the staff of the linguistics laboratory at the Ohio State University for the use of the laboratory facilities. Their kind assistance was a positive factor in the progress of this study. To my friend, Barb Connors, I offer my thanks for her help in typing and editing this manuscript.

The financial support of the Egyptian Government throughout my graduate study at the Ohio State University is gratefully acknowledged. Many thanks go to Prof. A. Rashwan and the staff of the Department of Biomedical Engineering at Cairo University for their continuing encouragement. Thanks also go to Prof. H. Weed and the foundation of Project HOPE for their kind assistance.

Finally, I would like to express my gratitude to my mother for her love, support, and encouragement without which I would have been unable to complete this work.
V I T A

July 9, 1954 ............. Born - Cairo, Egypt

1977 ...................... B. S., Electrical Eng.,
Cairo University,
Cairo, Egypt

1977-1980 .................. Teaching Assistant,
Cairo University,
Cairo, Egypt

1980-1982 .................. Project Hope Fellowship

1982 ...................... M. S., Biomedical Eng.,
The Ohio State University,
Columbus, Ohio

1982-1983 .................. Teaching Assistant, The Ohio
State University, Columbus,
Ohio

1984 ....................... Research Assistant, The Ohio
State University, Columbus,
Ohio

PUBLICATIONS

1. Application of the running spectrum analysis to study
the relaxation effects of the gastric electrical activity,

FIELDS OF STUDY

Major Fields : Physiological system modelling
Prof. H. R. Weed

Biological signal processing
Dr. R. M. Campbell

Minor Fields : Digital systems
Dr. J. Jagadeesh

Physiological systems
Dr. R. Hamlin

Computer Systems
Dr. M. T. Liu
# TABLE OF CONTENTS

| ACKNOWLEDGEMENTS                          | iii |
| VITA                                      | iv  |
| LIST OF TABLES                           | viii|
| LIST OF FIGURES                          | ix  |

## CHAPTER

### I. INTRODUCTION

1.1. Background and Motivation ............................................ 1
1.2. Goals and Objectives .............................................. 6
1.3. Organization .......................................................... 9

### II. SPEECH OF THE DEAF PERSONS

2.1. The Nature of Speech .................................................. 11
   2.1.1. Speech Production ................................................... 12
   2.1.2. Speech Perception ................................................... 15
   2.1.3. Levels of Speech Knowledge .......................................... 18
2.2. Characteristics of the Speech of Deaf Persons ......................... 20
   2.2.1. Intonation ........................................................... 22
   2.2.2. Formant Transitions ................................................ 24
   2.2.3. Timing Rates ........................................................ 25
   2.2.4. Intensity .............................................................. 26
   2.2.5. Articulation .......................................................... 26
2.3. Mechanisms of Deaf Speech Production .................................. 27
2.4. Summary ........................................................................... 28

### III. AUTOMATIC SPEECH RECOGNITION (ASR)

3.1. Isolated-Word Recognition .............................................. 31
   3.1.1. Feature Extraction .................................................... 32
   3.1.2. Pattern Matching ...................................................... 35
   3.1.3. Decision Rule .......................................................... 37
   3.1.4. System Performance .................................................... 38
3.2. The Problem of Segmentation ............................................ 39
   3.2.1. Definition of the Problem ............................................. 39
   3.2.2. Explicit and Implicit Segmentation .................................. 40
   3.2.3. Hybrid Recognizers .................................................... 42
   3.2.4. Passive and Active Models of ASR Systems ......................... 43
3.3. Automatic Speech Recognition for the Deaf ................................. 45
3.4. Automatic-Speaker Recognition .......................................... 46
   3.4.1. Measurements ............................................................ 46
   3.4.2. Classification ........................................................... 48
3.5. Linear Predictive Coding of Speech ..................................... 49
   3.5.1. Linear Prediction Model ............................................... 50
3.5.2. Speech Representation Using LPC ..... 52
3.5.3. Analysis Conditions ...................... 53
3.5.4. Speech Recognition Using LPC ............ 54

IV. METHODOLOGY ........................................ 56
4.1. Theory and Formulations ....................... 56
4.1.1. Identification
   of Deafness in Speech .......................... 57
4.1.1.1. Introduction ......................... 57
4.1.1.2. Speech Classification ................. 60
4.1.1.3. Generation
   of the Reference Vectors ..................... 62
4.1.2. Recognition
   of the Speech of Deaf Persons ............... 64
4.1.2.1. Introduction .......................... 64
4.1.2.2. The String Match Algorithm ........... 65
4.1.2.3. Computation
   of Word Match Distance ....................... 70
4.1.2.4. Generation
   of the Reference Patterns .................... 73
4.2. The AIRDS System ................................ 73
4.2.1. Speech Preprocessor ....................... 75
4.2.2. Speech Classifier ........................ 80
4.2.3. Speech Recognizer ......................... 82
4.2.4. The Trainer ............................. 85
4.3. Experimental Procedure ....................... 87
4.3.1. Subjects ................................ 87
4.3.2. Speech materials ......................... 87
4.3.3. Recording procedure ....................... 92
4.3.4. Analysis procedure ....................... 94
4.4. Computation of the Vocal Tract Area Function ............................................ 98

V. RESULTS .................................................. 101
5.1. Phonatory Results ............................ 101
5.1.1. Comparison
   of deaf and normal speech .................... 101
5.1.2. Statistical analysis ....................... 117
5.2. Classification Results ....................... 138
5.2.1. Classification of isolated words ........... 138
5.2.2. Classification
   of connected-word strings ..................... 141
5.2.3. Classification results
   for new utterances ............................ 142
5.2.4. Classification results
   for new speakers ............................. 143
5.3. Recognition Results .......................... 145
5.3.1. Isolated-word recognition ................... 145
5.3.2. Connected-word recognition ................. 150
5.3.3. Evaluation
   of the 2-step recognizer ..................... 155
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Deaf subjects</td>
<td>90</td>
</tr>
<tr>
<td>2. Speech materials</td>
<td>91</td>
</tr>
<tr>
<td>3. Fundamental frequency values for normal and deaf speech</td>
<td>109</td>
</tr>
<tr>
<td>4. Formant frequencies for normal and deaf speech</td>
<td>113</td>
</tr>
<tr>
<td>5. Energy levels and timing rates of normal and deaf speech</td>
<td>115</td>
</tr>
<tr>
<td>6. Acoustic measurements for normal and deaf speech</td>
<td>118</td>
</tr>
<tr>
<td>7. Correlation matrix for normal speech</td>
<td>119</td>
</tr>
<tr>
<td>8. Correlation matrix for deaf speech</td>
<td>120</td>
</tr>
<tr>
<td>9. Classification results using AFV</td>
<td>139</td>
</tr>
<tr>
<td>10. Classification results using LPV</td>
<td>140</td>
</tr>
<tr>
<td>11. Isolated-word recognition for normal speech</td>
<td>146</td>
</tr>
<tr>
<td>12. Isolated-word recognition for deaf speech</td>
<td>147</td>
</tr>
<tr>
<td>13. Summary of connected-word recognition</td>
<td>148</td>
</tr>
<tr>
<td>14. Connected-word recognition for normal speech</td>
<td>151</td>
</tr>
<tr>
<td>15. Connected-word recognition for deaf speech</td>
<td>152</td>
</tr>
<tr>
<td>16. Evaluation of the two-step recognizer</td>
<td>156</td>
</tr>
<tr>
<td>17. Summary of articulatory movements</td>
<td>168</td>
</tr>
<tr>
<td>18. Acoustic-articulatory correlation for normal speech</td>
<td>169</td>
</tr>
<tr>
<td>19. Acoustic-articulatory correlation for deaf speech</td>
<td>170</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The vocal tract</td>
<td>13</td>
</tr>
<tr>
<td>2. The hearing apparatus</td>
<td>19</td>
</tr>
<tr>
<td>3. A typical isolated-word recognition system</td>
<td>33</td>
</tr>
<tr>
<td>4. Word groupings in a 3-segment utterance</td>
<td>67</td>
</tr>
<tr>
<td>5. Blockdiagram of the AIRDS system</td>
<td>74</td>
</tr>
<tr>
<td>6. Flowchart of the Preprocessor</td>
<td>77</td>
</tr>
<tr>
<td>7. Flowchart of the Classifier</td>
<td>81</td>
</tr>
<tr>
<td>8. Flowchart of the Recognizer</td>
<td>84</td>
</tr>
<tr>
<td>9. Flowchart of training the Classifier</td>
<td>86</td>
</tr>
<tr>
<td>10. Flowchart of training the Recognizer</td>
<td>88</td>
</tr>
<tr>
<td>11. Blockdiagram of the recording system</td>
<td>93</td>
</tr>
<tr>
<td>12. Flowchart of the analysis procedure</td>
<td>96</td>
</tr>
<tr>
<td>13. Acoustic features of the utterance W11 produced by a normal-hearing speaker</td>
<td>102</td>
</tr>
<tr>
<td>14. Acoustic features of the utterance W11 produced by a deaf speaker</td>
<td>103</td>
</tr>
<tr>
<td>15. Acoustic features of the utterance W27 produced by a normal-hearing speaker</td>
<td>104</td>
</tr>
<tr>
<td>16. Acoustic features of the utterance W27 produced by a deaf speaker</td>
<td>105</td>
</tr>
<tr>
<td>17. Acoustic features of the utterance P7 produced by a normal-hearing speaker</td>
<td>106</td>
</tr>
<tr>
<td>18. Acoustic features of the utterance P7 produced by a deaf speaker</td>
<td>107</td>
</tr>
<tr>
<td>19. Intonation contours for the utterance P7</td>
<td>110</td>
</tr>
<tr>
<td>20. Transitions of the second formant for the utterance P7</td>
<td>114</td>
</tr>
</tbody>
</table>
21. Histogram of the fundamental frequency for normal speech ........................................ 122
22. Histogram of the fundamental frequency for deaf speech ........................................ 123
23. Histogram of the second formant frequency for normal speech ................................ 125
24. Histogram of the second formant frequency for deaf speech ................................... 126
25. Histogram of the first reflection coefficient for normal speech ................................. 127
26. Histogram of the first reflection coefficient for deaf speech .................................... 128
27. 2D scatter plot of the first and second formant frequencies .................................... 129
28. 2D scatter plot of the first and second reflection coefficients .................................... 130
29. 2D scatter plot of the fundamental frequency and second formant frequency ........... 132
30. 2D scatter plot of the fundamental frequency and the first reflection coefficient ........... 133
31. 2D scatter plot of the first and second formant frequencies using different sets of utterances for normal and deaf speech ........................................ 134
32. 2D scatter plot of the first and second reflection coefficients using different sets of utterances for normal and deaf speech ........................................ 135
33. 2D scatter plot of the fundamental frequency and second formant frequency using different sets of utterances for normal and deaf speech .......................... 136
34. 2D scatter plot of the fundamental frequency and first reflection coefficient using different sets of utterances for normal and deaf speech .......................... 137
35. Classification results for normal speech ........... 143
36. Classification results for deaf speech .......... 144
37. The vocal tract area at the tongue position for the utterance W11 ............................. 159
38. The vocal tract area at the tongue position for the utterance W27 ............................. 160
39. The vocal tract area at the tongue position for the utterance P7 ............................. 161
40. The vocal tract area at the larynx position for the utterance W11 ............................. 163
41. The vocal tract area at the larynx position for the utterance W27 ............................. 164
42. The vocal tract area at the larynx position for the utterance P7 ............................. 165
CHAPTER I
INTRODUCTION

1.1. Background and Motivation:

Speech is the most convenient and efficient form of communication between human beings. Unfortunately, few deaf people ever attain a speech quality that is adequate for normal conversation (the term deaf or hearing impaired is used in this study to refer to the congenitally profoundly deaf persons who lost their hearing in the first few years of life and whose hearing loss is greater than 90 dB). For many years, it was hoped that this handicap could be overcome through the use of speech processing aids which include speech training aids and speech perception aids.

Speech training aids typically operate by extracting information from the speech signal and presenting that information back to the deaf user in either visual or tactual form. This information, which can be contrasted to patterns formed when the sound is produced correctly by a normal hearing speaker, provide the deaf individual with visual or tactual feedback they may use in learning to speak in a manner similar to the way auditory feedback helps people with normal hearing. The assumption being
that the speech production-perception feedback loop is closed through another sense modality.

Thus far, speech training aids have shown a limited degree of success as speech improvement aids (Pickett, 1969; Levitt, 1973; Nickerson, 1976; Rothman, 1976; and Waldron, 1982). Nickerson et. al (1976) reported that measurable improvements in overall intelligibility of unrehearsed speech (utterances that are not used during training) were not demonstrated for most deaf subjects. Rothman (1976) stated that speech training could be faulty, and that it might intensify the problems of deaf speech. Waldron (1982) reported that, despite the intensive rehabilitation work, the congenitally deaf children retained a distinctive deaf speech quality, exhibiting different temporal and muscular control.

Speech perception aids typically operate by converting the speech signal into a feature representation, comparing this with templates of known utterances, and selecting the best match as the input utterance. The use of these aids to overcome the handicap of deaf persons, that is to recognize deaf speech, has not been actually investigated. The reason is that their development has been pending on the progress in speech recognition techniques as well as understanding
the acoustic and perceptual characteristics of the speech of deaf persons.

Research in automatic speech recognition has advanced to the stage where several commercial systems are now available. These systems can reliably recognize words from a vocabulary of up to several hundred words and phrases with accuracies greater than 95 percent. Although most of these systems recognize isolated words and they are speaker dependent, there is at least one speaker-independent isolated-word recognizer and another speaker-trained connected-word recognizer (Rabiner et al., 1981). Other systems are used for speaker identification and verification purposes (Doddington, 1985). It should be noted that these systems are all applied to speech produced by normal hearing speakers.

On the other hand, advances have been made within the last decade in studying the speech patterns of the hearing impaired persons. The speech of the deaf persons has been described as flat, monotonous with insufficient intonation. It is found to be slow and inefficient due to prolongation of individual sounds or pausing for breath more frequently than normal (Levitt, 1973). Research in this field has shown that although deaf individuals often fail to follow typical pattern of normal speech, the
deviations in their speech show systematic patterns and do not generally occur in a random way. These studies concluded that the deviations shown do not appear to be simple errors of speech production, but reflect instead a different type of coding structure (Osberger et. al, 1982) and perhaps a different sense modality in capturing speech (Waldron, 1982).

Yet, the double communication handicap of speaking and hearing that the deaf persons face has not been overcome. Speech training aids could be helpful; however, and despite the extensive training, the resulting speech still has a deaf-speech quality which is not adequate for normal conversation. When communicating with other persons, whether in person or over a telephone line, this deviant speech quality leads to miscommunication. In addition, and because normal speech is often un-intelligible to deaf persons, speech training by itself is not sufficient to overcome this communication barrier.

Other methods such the TDDs (Telecommunication Devices for the Deaf) and sign language have been used to tackle this problem. The TDDs are used to facilitate communication over a telephone line where a typewriter with a visual display is used at both ends of the line.
Finger spelling (sign language) is used to achieve person-to-person communication where hand and finger movements are used to represent speech utterances. Lip-reading is another, but difficult method in person-to-person communication.

These methods do not use speech as the means of communication, hence, they prove to be neither efficient nor convenient. This is particularly so, if normal hearing individuals, who rely heavily on speech, are communicating with the hearing impaired individuals. The TDDs require the person on the other end of the telephone line to have a similar device. They also require efficient typing and writing skills which are difficult at best for deaf persons. In addition, Signing and Lipreading require special skills and extensive training.

In light of the above discussion, the following points can be made:

First, speech training of the deaf persons has shown to be of limited success since deaf and normal hearing individuals do not follow the same rules for speech production or the same sense modality for speech perception. As a result, deaf persons develop speech with deaf quality which is not adequate for normal conversation.
Second, the need for an alternative means of communication for the deaf persons is evident and justified. Available methods such as Finger spelling, Lipreading, or the TDDs are neither efficient nor convenient.

Finally, and due to the advances made in recognizing speech by machines as well as understanding the acoustic and perceptual characteristics of the speech of the deaf, the use of speech perception aids to overcome the handicap of the deaf seems promising and convincing. Speech perception aids could be used by deaf individuals to recognize normal speech, or by normal hearing persons to understand deaf speech. Successful automatic recognition of deaf speech along with a visual display could be used to develop a communication aid for the deaf persons. This aid could be used to communicate with other individuals or to operate voice controlled machines around home and work.

1.2. Goals and Objectives:

The goal of the present work is to investigate the possibility of using the techniques of automatic speech recognition to develop a computer system for Automatic Identification and Recognition of Deaf Speech. This will be referred to as the AIRDS system.
To achieve this goal, three objectives are to be identified:

1) The first objective is to study the speech characteristics of the deaf persons and its deviations from normal speech. This will help establish a set of measurements which could be effectively used to distinguish between the two kinds of speech. An attempt to study the articulatory movements during speech production by the deaf persons will also be made.

2) The second objective is to identify deafness in speech, that is, to decide whether a given speech utterance can be classified as normal or deaf speech.

3) The final objective is to recognize the speech message contained in the input utterance whether it is classified as normal or deaf speech.

Identification of deafness in speech will be treated as a speaker-recognition problem in which only two speakers (or actually two classes of speakers) are recognized based on measurements made on the speech signal. The measurements used in this study will represent four speaker dependent features; namely: intonation contours, formant transitions, timing and speaking rates, and energy contours. The use of other
speech features such as the LPC (linear predictive coding) parameters and the VTAF (vocal tract area function) for speech classification will also be investigated.

Due to the nature of deaf speech, where frequent and lengthy inter-, and intra-word pauses are inserted a two-step hybrid recognizer will be employed in the AIRDS system. In this recognizer, LPC parameters will be used for speech representation and the nearest neighborhood rule will be used for determination of pattern similarity. Reference patterns will be generated by combining (time warping) several repetitions of a word to form a reference template of that word.

The AIRDS system will be implemented on a PDP-11/23 minicomputer with the Interactive Laboratory System (ILS) software. A total of 468 utterances spoken by two deaf, male speakers and two normal hearing male speakers will be used to evaluate the system. These utterances represent 8 phrases and 31 words which are a subset of an artificial recognition language previously designed and used for airline information and reservation tasks.

The impact of the AIRDS system will be seen in developing a new communication aid for the deaf persons. This aid is envisioned as a device that accepts the speech
of deaf speakers as well as normal hearing speakers, identifies its kind, recognizes it, and finally produces it back for communication. In addition, this system will make voice-controlled machines of the future equally accessible to the deaf speakers. Further, this system could be modified in ways to study accent differences so that voice-recognition machines become more general and require less training by individual users.

1.3. Organization:

A literature review of speech production and perception models as well as the characteristics of the speech of deaf persons and their deviations from normal speech is presented in chapter 2.

In chapter 3, the basic concepts of automatic speech recognition along with related literature are discussed. First, the simple isolated-word recognition systems are introduced. Second, the problem of segmentation in connected speech is presented. Next, the concepts of automatic speaker recognition are described. Finally, a brief description of the method of linear prediction coding of speech is given.

In chapter 4, the AIRDS system for automatic identification and recognition of deaf speech is
presented. First, the method proposed to identify and recognize deaf speech is described. Description of the basic components of the system; namely: the preprocessor, classifier, recognizer, and trainer, is given. Finally, in this chapter, the experimental procedure used to test and evaluate the system is described.

The results are discussed in chapter 5. A summary, conclusions, and future directions of this work are presented in chapter 6.
CHAPTER II

SPEECH OF DEAF PERSONS

In order to apply the techniques of automatic speech recognition to identify and recognize deaf speech, it is essential to understand the characteristics of deaf speech and its deviations from normal speech. In this chapter, a review of the speech characteristics of the deaf persons is provided. First, a summary of the nature of speech along with the models of normal speech production and perception are given. Then, the characteristics of deaf speech and their deviations from normal speech are described. Finally, a summary of these characteristics is given.

2.1. The Nature of Speech:

Speech is a form of communication between human beings which involves the generation and reception of a complex acoustic signal. The generation and reception of speech, which may be thought of as coding operations, take place over a hierarchy of processing levels. At the highest level, thoughts to be communicated are formed. At lower levels, these thoughts are encoded in the form of words. At the lowest level, these words are encoded in the form of neural and articulatory activities to produce
the speech sounds.

Reciprocal processes are used by the listener to receive these thoughts. At the lowest level, the spoken words are perceived as consisting of successions of elemental sound units called phonemes. At higher levels, these sound units are decoded in the form of neural patterns. At the highest levels, these patterns are decoded to receive these thoughts (Hyde, 1979). These coding operations to produce and perceive speech are discussed next.

2.1.1. Speech Production:

The so-called Source-Filter model of speech production is presented here. This is also referred to as the acoustical theory of speech production which was first presented by Fant et al. (1960) for consonant and vowel production, and Stevens et al. (1961) for vowel production. In this model, as described by Hill (1980), speech is generated and received as a continuously time-varying pressure waveform. This wave originates from voluntary movements of the structures shown in Figure 2. Forcing air through the vocal folds and causing them to vibrate, or forcing air through a relatively narrow constriction in the vocal tract, acts as a source of
Figure 1. The human articulatory system.
energy that may excite in the tube-like vocal tract which extends from the folds to the lips. A second resonance cavity, leading to the nostrils, may be connected in parallel by opening the velum.

The energy due to the air flowing through the vibrating vocal folds is termed as voiced energy, and the state is called voicing resulting in voiced sounds. The rate at which the vocal folds vibrate is dependent upon the air pressure in the trachea and the physiological adjustments in the vocal folds which include changes in their length, thickness, and tension. The greater the tension the higher the perceived pitch (or the acoustically measured fundamental frequency) of the voice produced.

The other kind of energy, that is, due to air flowing through a constriction in the vocal tract, is called fricative leading to unvoiced fricative sounds. The two kinds of energy may be mixed resulting in voiced-fricative sounds. A third source of energy may be due to a build up of air in the mouth and then suddenly release it leading to plosive sounds.

The energy supply constitutes the source in the source-filter model of speech production. The source
energy is distributed through the frequency spectrum, falling off with increasing frequency at a fairly uniform rate. In passing through the resonances of the tract, the distribution of energy is modified with peaks being produced at or near the resonant frequencies of the tract.

These peaks are called formants. The resonant frequencies and hence the frequencies of the formants vary according to the shape of the vocal tract. Although there is a series of formants, decreasing in amplitude with increasing frequencies, only the three formants lowest in frequency are necessary for good speech intelligibility. The connection of the nasal cavity further modifies the frequency distribution. Thus, the overall effect of the vocal tract, mouth, and nasal cavity (if connected) is to act as a filter on the source energy (Hill, 1980).

2.1.2. **Speech Perception**

Speech perception stands for the sensation and interpretation of the incoming speech signal based on experience acquired by auditory feedback. There are two groups of theories on how speech perception is accomplished. In one group, the listener is relatively
passive and the process of speech perception as primarily sensory. The message is sensed, filtered, and mapped directly onto the acoustic phonetic features of the language (phonemes, syllables, or words). These passive theories of speech perception, presented by Fant (1967) and Morton et al (1967), emphasize the sensory filtering mechanisms of the listener while ignore the role of speech production to a minor secondary place used only when perceptual tasks are difficult (Borden et al, 1981).

The other group of theories of speech perception views the listener as more active and postulates that the process of speech perception involves some aspects of speech production. The sounds are sensed, analyzed for their phonetic features by reference to how such sounds are produced, and thus, recognized. This group of active theories includes the Motor Theory of speech perception by Liberman et al (1967) and the Analysis-by-Synthesis Theory of speech perception by Stevens et al (1967).

The Motor Theory and the Analysis-by-Synthesis Theory of speech perception are similar in that they relate and refer speech perception to speech production. However, they differ in that the reference to speech production is more articulatory in the first, while it is more acoustic and relies on a system of matching in the second.
In the Motor Theory, the listener's articulatory knowledge is used to mediate between the acoustic signal and the phonetic or phonemic information since there is a lack of correspondence between the acoustic event and the separate phonemes (Borden, 1981). This theory was revised to accommodate recent findings and to relate the assumptions of the theory to those that might be made about other perceptual modes (Liberman et al, 1985).

In the Analysis-by-Synthesis Theory, on the other hand, the listener receives the input auditory pattern and analyzes it by generating a pattern of his own, and finally perceives it if the two patterns match. The main notion about the Analysis-by-Synthesis model of speech perception is that the listener can synthesize patterns based on experience acquired from speech production, and match these internally generated patterns with the unknown patterns to analyze and eventually perceive them. In chapter 3, it will be shown that speech recognition by machines is done in a similar manner. For this reason, this model is described here in some details.

According to the Analysis-by-Synthesis model, human perception of speech involves a frequency analysis performed on the speech signal by the peripheral auditory
system (the ear) shown in Figure 2. The results of this
frequency analysis are then transformed into neural pulses
that are interpreted in the brain as speech sounds.

This transformation is performed by the basilar
membrane of the inner ear. The basilar membrane is
frequency selective such that the pattern of the neural
impulses transmitted from the ear to the brain resemble
the spectrum of the input signal. This resulting pattern
is temporarily stored in the cortex while being compared
with templates internally generated based on experience
acquired from monitoring speech. Once a match occurs,
the input speech signal is recognized (Dew et. al, 1977).

Finally, it should be mentioned that the TRACE model
of speech perception by McClelland and Elman (1986) is
another active theory of speech perception. This model
uses lexical information to identify phonemes and words
and to influence processing of non-word utterances.

2.1.3. Levels of Speech Knowledge:

It should be mentioned that speech production and
perception as described before are at the lowest
(acoustic) level of speech knowledge, that is, the
production and perception of individual speech sounds such
as phonemes, syllables, or words. However, speech sounds
Figure 2. A simplified diagram of the auditory neural pathways.
are rarely produced in isolation; they overlap and influence one another as a result of their production.

For perception, this means that speech sounds often are not discrete or separable. Therefore, the listener's must use higher levels of speech knowledge such as the syntactic, semantic, and pragmatic levels to aid the decoding of the input auditory signal. Human perception of speech is thus likely to be an active process in which higher levels of speech knowledge are used to guide the lower levels in decoding the input utterance (De Mori, 1983).

Speech decoding at these higher levels is a specialized and lateralized function in the brain. Yet, this is poorly understood when compared to the understanding of the acoustics of sound generation by the vocal tract and the physics of sound analysis by the peripheral ear. As a result, automatic speech recognition by machines can now emulate the lower (peripheral) levels well, but it is far away from the higher (brain) levels (Flanagan, 1982). This issue will be addressed again in chapter 3.
2.2. **Characteristics of the Speech of Deaf Persons**

Speech is a motor behavior that must be learned. It is developed, controlled, and maintained by the acoustic feedback of the hearing mechanism and by the kinesthetic feedback of the speech musculature. Information from these senses is organized and used by the central nervous system to direct speech functions. Impairment of either control mechanism usually degrades the performance of the vocal tract and hence the quality of the resulting speech (Flanagan, 1972). This is the case of individuals with congenital or profound hearing loss where due to the lack of auditory feedback, normal development of speech is disrupted. Consequently, most hearing impaired children must be taught the speech skills that normal hearing children readily acquire during the first few years of life. Although some deaf children may develop intelligible speech through training, many do not. Despite the intensive training, children often retain a distinctive deaf speech quality, exhibiting a lack of temporal and muscular control.

Research in this field has shown that the hearing impaired persons often fail to follow rules typical to those of normal speech; however, the deviations in their speech show systematic patterns and do not generally occur in a random way. Osberger et. al (1982) stated that
these deviations do not appear to be simple errors of speech production, but reflect instead a different type of coding structure in producing speech.

Waldron (1982) stated that these deviations are due to the use of different sense (coding) modalities in the development of speech of deaf persons. She hypothesized that the deaf individuals use tachemic code to acquire and achieve their speech as compared to the phonemic code used by the normal hearing individuals.

King et. al (1982) classified these deviations (referred to as speech "problems" of the deaf persons) into: breath control problems (inadequate intake, excess loss of air whilst speaking, and use of non-pulmonic air stream mechanisms for speaking), voice production problems (weak phonation, creaky or breathy voice, and poor coordination of voice with co-articulation), timing problems (breathing between words or syllables, abnormally slow rate of utterance, prolonged sounds, and lack of stress—short vs long syllables and extra or missing syllables), intonation problems (flat pitch, and no use of pitch marking of phrases or important words within phrases), and speech sounds problems (failure to produce acoustic cues to speech sounds contrasts).
A comprehensive description of the characteristics of deaf speech and their deviation from normal speech is given by Osberger et. al (1982), and Nickerson (1975). A brief description of these characteristics is given here.

2.2.1. Intonation:

Poor control of the vocal folds results in poor intonation with insufficient or excessive variability. Intonation is the perceived pattern of change in the fundamental frequency along an utterance. The declination (gradual reduction) as well as the terminal fall (sharp drop in the fundamental frequency at the end of the utterance) are absent in the speech of deaf persons (Osberger et. al, 1982). Levitt (1972) described the speech of the hearing impaired speakers as being frequently flat and monotonous.

For this, intonation contours are very effective to discriminate between different speakers. Levitt (1972) stated that the shape of the fundamental frequency contours is a key variable in the design of speech training aids for the deaf persons. Monsen (1979) used intonation contours as an important characteristic in separating the better from the poorer hearing impaired speakers.
For a simple declarative sentence spoken by a profoundly hearing impaired speaker, a variation of about 25 Hz was found in the fundamental frequency. This was judged as insufficient variability compared to a change of about 100 Hz obtained when the same utterance was spoken by a normal hearing speaker. However, other hearing impaired speakers produced an excessive variation of about 200 Hz in producing the same utterance (Osberger et. al, 1982).

For normal hearing adults, the fundamental frequency ranges from 100 Hz to 175 Hz for males, and from 175 Hz to 250 Hz for females (Nickerson, 1975). In the speech of deaf males, the average value of the fundamental frequency is about 50 Hz lower than that of normal hearing males. However, the average value of the fundamental frequency for deaf female speakers has been reported as about 75 Hz higher in some cases and about 50 Hz lower in others than that for normal hearing female speakers (Osberger et. al, 1982). Due to this uncertainty, it was decided that only male speakers will be used in the present study.

2.2.2. Formant Transitions:

In the speech of the deaf persons, and due to poor articulation, transition of all formants in general and the second in specific are reduced in both time and
frequency when compared to normal speech (Osberger et. al., 1982).

Monsen (1976) reported that the second formant for all vowels remained around 1800 Hz rather than varying as different vowels were articulated in the speech of deaf speakers.

Rothman (1976) also reported that the slopes of the formant transitions were found fairly flat when either a rising or falling pattern was dictated in the speech of deaf persons. For the deaf speakers, these slopes varied from 160 Hz to 300 Hz during the first 120 milliseconds. For normal speakers, these values were from 380 Hz to 700 Hz during the same period when producing the same utterances (Monsen, 1976).

2.2.3. Timing Rates:

Poor timing control in the speech of the deaf persons results in prolongation of speech segments and insertion of frequently and lengthy pauses often at syntactically inappropriate boundaries. Nickerson (1975) stated that deaf speakers often fail to insert pauses at certain phrase boundaries, or they insert pauses at inappropriate places in a sentence. More and longer inter-, and intra-word pauses were observed in the speech of deaf
persons as compared to normal speech (Stevens et. al, 1978).

Deaf speakers as a rule tend to produce sounds inefficiently, drawing out the length of individual sounds and pausing for breath more frequently than normal speakers (Levitt, 1972; and Nickerson, 1975). On the average, deaf speakers take about 1.5 to 3 times longer to produce the same utterance as do normal hearing speakers (Nickerson, 1975; and Osberger et. al, 1982).

2.2.4. Intensity:

Improper adjustment of the vocal folds in the speech of the deaf persons results in acoustic signal with enhanced energy in the low frequency range and deficient energy in the high frequency range (Stevens et. al, 1978).

Nickerson (1975) reported that in the speech of the deaf speakers, voicing may be too loud or too soft and intensity may vary erratically. In general, the energy level and range are lower in the speech of the deaf persons than for normal hearing speakers (Osberger et. al, 1982).
2.2.5. **Articulation**

Typical articulatory errors in the speech of the deaf persons include voicing errors, substitution errors, and omission errors. Voicing errors are common in the speech of the deaf persons and they often involve voiced for voiceless substitution. Substitution errors involve the substitution of one phoneme for another; frequently, the substitution to a phoneme with a similar place of articulation particularly those produced in the middle or the back of the mouth. Omission errors most often involve consonants, particularly those in the word-final position while omission of vowels is infrequent and usually does not occur unless the entire syllable is omitted (Osberger et. al, 1982; and Nickerson, 1975).

2.3. **Mechanisms of Deaf Speech Production**

There are few studies that describe the physiological correlates of the deviations of deaf speech from normal speech. These physiological studies include studies on the respiratory patterns of deaf speakers (Whitehead, 1982; and McGarr et. al, 1982), studies to examine the phonatory mechanisms in deaf speakers (Monsen et. al, 1979; and Metz et. al, 1982), and studies of articulation in the deaf speakers (McGarr, 1980; and Zimmerman et. al, 1981).
Osberger et al. (1982) discussed these studies and concluded that some of the production mechanisms responsible for the perceptual and acoustic distortions in deaf speech could be: posture characteristics and interarticulatory coordination. Improper posture characteristics include: improper respiratory control, glottal abduction/adduction, tongue position and range of movement. Difficulties in interarticulatory coordination of the complex activities of respiration, phonation, and articulation by deaf speakers are evidenced by poor control and poor coordination of articulatory gestures, both at the laryngeal and supralaryngeal levels of production.

Metz et al. (1984) studied the laryngeal articulatory gestures in deaf speakers. They suggested that deaf persons may exhibit at least two fundamental laryngeal dysfunctions during speech production. One is the inability to control certain laryngeal muscles responsible for maintaining periodic vocal fold oscillation and the second is the inappropriate positioning of certain laryngeal structures during speech production.

2.4. Summary:

To summarize, the deviations in the speech of the deaf persons from normal speech can be cited as:
1. Intonation is either flat or excessive. The average value of the fundamental frequency is about 50 Hz lower for deaf males than that for normal hearing males. In deaf speech, the average range of variation in intonation is about 25 Hz when described as insufficient and about 200 Hz when described as excessive as compared to a range of about 100 Hz obtained in normal speech.

2. Transition of all formants in general and the second in specific are reduced in both time and frequency. The average range of change in the second formant in deaf speech is about one half that of normal speech.

3. More and longer inter- and intra-word pauses are inserted in the speech of deaf speakers. Also, speech segments are about 2 to 3 times longer in deaf speech as compared to normal speech.

4. The energy level and range are lower in the speech of deaf speakers when compared to normal hearing speakers.

The above four characteristics; namely: intonation contours, formant transitions, timing rates, and intensity contours offer a set of speaker-dependent features that can be used effectively to distinguish between deaf and normal speech. In chapter 4, it will be shown how to use
these features to identify deafness in unknown speech, that is, to classify a given speech utterance into either normal or deaf speech.

Other deviations from normal speech are also common in the speech of deaf persons. These include omission errors: omission of word-initial or word-final consonants, substitution errors: substitution of one phoneme for another, and voicing errors: confusion of the voiced-voiceless distinction. Further, abnormal voice characteristics such as harshness, hyper- and hypo-nasality may also be present in the speech of the deaf persons. However, and since quantitative measurements of these characteristics are difficult to obtain, they will not be considered in the present study.
CHAPTER III

AUTOMATIC SPEECH RECOGNITION

Automatic Speech Recognition (ASR) is a term used to refer to understanding speech by machines. Advances have been made in this field to the state where it is now possible to communicate reliably with computers by speaking to it in a disciplined manner using a vocabulary of moderate size (Rabiner et al., 1981). The general principle of automatic speech recognition involves comparing a feature representation of an input utterance with that of the reference template of each word in the vocabulary, and selecting the best match using some distance rule. The complexity of such a recognition system is determined by three factors; namely: the vocabulary size, whether or not the system can accommodate any speaker or only those who have trained the system, and whether the input speech is individual isolated words, a string of connected words, or continuous speech (Reddy, 1976).

In this chapter, the simple case of a small-vocabulary speaker-trained isolated-word recognition system will be considered first. Second, the problem of segmentation in connected speech will be
addressed. Third, future applications of automatic speech recognition for the benefit of the deaf persons will be introduced. Next, the basic concepts of automatic speaker recognition will be presented. Finally, a brief description of the method of Linear Predictive Coding of speech will be given.

3.1. **Isolated-Word Recognition Systems**

In an isolated-word recognition system, typically a set of small number of words produced in isolation by known speakers are recognized. In this case, the problem of automatic speech recognition is divided into three parts; namely: feature extraction, pattern similarity determination, and decision rule classification (Rabiner et. al, 1981). A block diagram of a typical isolated word recognition system is shown in Figure 3.

3.1.1. **Feature Extraction**

Feature extraction is basically a data reduction technique where a large number of data points (speech samples) are transformed into a smaller set of features that faithfully represent and describe the speech signal. For speech recognition systems, two feature sets have been frequently used. One is formed from energy levels and zero-crossing rates in selected frequency bands, and the second is formed from linear predictive coding (LPC)
Figure 3. A typical isolated-word recognition system.
coefficients.

The first feature set is based on the filter-bank analysis which models the auditory frequency analysis performed by the basilar membrane of the ear (De Mori, 1983). In this model, the speech signal is passed through a bank of band-pass filters covering the speech band (from about 100 Hz to about 5000 Hz). The output of each filter is passed through a detector or full wave rectifier, and then low-pass filtered to give a signal which is proportional to the energy of the speech signal in the band represented by that filter.

For a given speech segment (typically a time frame of about 20 to 50 milliseconds), the parallel outputs of, say K filters, define a feature vector of K elements. The time course of this vector, that is for consecutive frames, defines a feature pattern of the input utterance. A zero-crossing counter may be added at the output of each filter giving a measure of the formant frequency in that band, resulting in a feature vector of 2K elements (Rabiner et. al, 1981).

The second feature set that is frequently used in automatic speech recognition systems is based on Linear Predictive Coding analysis. In LPC analysis, a given
speech sample can be approximated by a linear combination of the past values of the signal. By minimizing the sum of the squared differences between the actual and the predicted samples over a finite interval, a unique set of predictor coefficients can be determined (Markel et. al, 1976). These coefficients constitute a feature vector which represents the speech signal in that interval. The time course of this vector, for consecutive intervals, defines a pattern of the input utterance.

Given a set of predictor coefficients, other parametric representations of speech such as spectrum, pitch, formants, and vocal tract area function can be obtained (Atal, 1976). The method of LPC was shown to be closely related to the basic model of speech production in which the speech signal is modeled as the output of a linear, time-varying system excited by either quasi-periodic pulses for voiced sounds or random noise for unvoiced sounds (Markel et. al, 1974). The method of LPC will be used in the present system, and the procedure of obtaining an LPC-based feature vector is described in section 3.5.

3.1.2. Pattern Matching

The second part in an isolated-word recognition system is to determine the similarity between test and
reference patterns. This involves time alignment and distance computation.

Because speaking rates vary greatly, the time scales of a test and a reference patterns are seldom aligned in time even for utterances of the same word and produced by the same speaker. Approximate time alignment can be achieved by synchronizing the beginning and end of the two utterances. This is a simple linear compression or expansion of one time scale to match the other.

In other cases a nonlinear time-warping is used to compensate for local compression or expansion of the time scale. The method of dynamic time warping (DTW) is used in such cases to map different parts (speech events) in the test pattern onto the corresponding parts in the reference pattern. The criterion for correspondence is that some local distance, which represents a measure of dissimilarity between a frame in the test pattern and another in the reference pattern, is minimized.

To compute the local distance in speech recognition systems, measures such as the Euclidean, Covariance Weighting, and LPC-log have been used (Rabiner et. al, 1981). The word distance, which represents a measure of dissimilarity between the test and reference patterns, is
computed as the sum of the local distances for all frames in the test pattern. In chapter 4, it will be shown how to use the DTW method to compute the word distance using the Euclidean measure for local distance computation.

3.1.3. Decision Rule:

The last part in automatic speech recognition systems is the use of some decision rule to choose which reference pattern closely matches the input test pattern. In most practical systems, only two decision rules have been used. These are the nearest neighborhood rule and the K-nearest neighborhood rule (Rabiner et al, 1981). The nearest neighborhood rule chooses the reference pattern with the minimum word distance to the test pattern as the recognized word.

The K-nearest neighborhood rule is applied in recognition systems where each word in the vocabulary is represented by two or more reference patterns; for example, in speaker-independent recognition systems. For each word in the vocabulary, K distances are computed when compared with the input test pattern. These distances are then averaged, and the reference pattern with the minimum average distance is chosen as the recognized word. The nearest neighborhood rule will be used in the present system.
3.1.4. System Performance:

The above three parts; namely: feature extraction, pattern matching, and decision rule, constitute the basic design of any isolated-word recognition system. Such a system has two distinct modes of operations. The first is training, where the reference patterns of the vocabulary are created and stored in the system memory which is also shown in Figure 3. A reference pattern of a word is created by combining or averaging several tokens (repetitions) of that word. The method of dynamic time warping (DTW) is again used here to compensate for differences in speaking rates.

The second mode of operation in isolated-word recognition systems is testing, where the input utterance is matched against all reference templates, and eventually recognized. The accuracy with which the recognition task is performed, known as the overall system performance is the ultimate criterion in the design of that system. However, other factors such as the computation time, storage, and ease of implementation are also important (Rabiner et. al, 1981). As the vocabulary size increases or as in the case of independent-speaker recognition systems, the number of reference patterns increases and better search techniques are needed to achieve the recognition task efficiently.
Rabiner et. al (1980) used a two-pass approach to recognize similar words. The first pass is used to find an equivalent class of similar words, while the second pass is used to find the best match of the input word in the selected class. Kaneko et. al (1983) used a two-step hierarchical approach for large vocabulary recognition systems. In the first step, only two features were used to find a list of best match candidates. In the second step, all features are used to find the best match word in that list.

3.2. The Problem of Segmentation

3.2.1. Definition of the Problem

The spoken input in isolated-word recognition systems is individual words produced in isolation, whereas the spoken input in continuous speech recognition is a string of articulated words. A special case is when the sequence of words are produced in a disciplined manner and possibly with short pauses inserted between the words in that string. This case is referred to as connected speech recognition.

The philosophy of isolated-word recognition, based on whole word template matching as described above, can be extended to recognize an utterance of connected words or continuous speech. However, this requires determining
the beginning and ending of each word in the utterance which is known as speech segmentation. Speech segmentation is the key problem in connected-word and continuous speech recognition systems, and it is more difficult and unreliable in the latter case due to co-articulation effects at word boundaries.

3.2.2. Explicit and Implicit Segmentation:

Two approaches for word segmentation have been employed in connected speech recognition systems in which a set of isolated-word templates are used as reference patterns. In the first approach, the word boundaries are explicitly determined prior to and independent of the recognition task (Sambur et. al, 1976; and Zeliniski et. al, 1983). This is done primarily by marking drops in the energy of the input speech signal.

The second approach is completely different and is based on a recognition procedure where no explicit segmentation is done before recognition, but the word boundaries are implicitly estimated during recognition (Sakoe, 1979; Rabiner et, 1980; and Myers et. al, 1981). A set of concatenated reference patterns are matched against the the incoming utterance (string of connected words). The concatenation which yields the best match determines the recognition results and implicitly the word
boundaries.

With the second approach, higher recognition accuracy is expected since the possibility of inaccurate segmentation is excluded. However, this approach requires an enormous amount of computation time since all possible concatenations of reference patterns must be tested. This is particularly so in large-vocabulary systems, speaker-independent systems, and other recognition systems in which a large number of reference templates are used. In addition, concatenating isolated-word templates still does not account for coarticulation effects that occur at word boundaries. This co-articulation occurs to a limited degree in connected speech, but mostly in continuous speech.

Several efficient implicit-segmentation algorithms have been developed to reduce the amount of computations involved in connected word recognition. Among these are the one-stage dynamic programming algorithm (Ney, 1984), the level-building algorithm (Myers et. al, 1981), and the two-level building algorithm (Sakoe et. al, 1979). The syntax-directed level building algorithm (Myers et. al, 1982) is another example. This latter algorithm differs from the former three in that the syntax, which represents high level of speech knowledge, is used to guide the
recognition task.

3.2.3. Hybrid Recognizers

A hybrid recognizer which combines features from the two segmentation approaches just described was suggested by Zeliniski et. al (1983) and used by Brassard (1985) for connected and continuous speech recognition. Brassard's recognizer first segment the unknown utterance into pseudo syllables, and treats this a priori segmentation as initial boundary hypothesis. Then, and during recognition phase, it time-warps the incoming part of the utterance to a selected template; and a similarity derived boundary is found. When the new and the initial boundaries do not coincide, the recognizer considers both and pursues them in parallel untill further recognition shows which one to select. The results obtained show a reduction in the error rate by a factor of three with no considerable cost in computation time.

In the present study, a hybrid recognizer is used to recognize the speech of the deaf. This recognizer is based on the string match algorithm used by Welch et. al (1980) to improve data entry rate in isolated-word recognition systems.
3.2.4. Passive and Active Models of ASR Systems:

Automatic recognition of isolated words as described above simulates the passive model of human perception of speech discussed in chapter 2. The microphone, shown in Figure 3, roughly corresponds to the outer ear. The part of feature extraction corresponds to the inner ear and the auditory nerve, and pattern matching corresponds to the process of matching done in the cortex which acts as a memory of templates (De Mori, 1983; White, 1979).

Unfortunately, recognition of connected and continuous speech cannot be handled by simply matching each word in the utterance against the reference templates. Finding the boundaries between words in connected and continuous speech is not an easy task. In addition, the pronunciation of individual words changes when the words are spoken continuously. For example, a whole syllable may be dropped or "swallowed" at word boundaries (Barr et. al, 1981). Further, it is important in some situations to understand not only the meaning of each word, but also the intent of the message contained in the utterance (Atal, 1976).

It was shown in chapter 2 that human perception of speech is an active process in which expectation as well as cognition are used to aid the decoding. All facets of
the listener's knowledge such as syntax, semantics, pragmatics, prosody, as well as phonology are used to understand what is being said. An active model of automatic speech recognition requires the representation of this knowledge at all levels (De Mori, 1983).

Active models of automatic speech recognition are usually referred to as Speech Understanding Systems (SUS) in which knowledge about syllables, words, and sentences; about rules of conversations; and about the subject under discussion is often used in recognition. In speech understanding systems, the techniques of Artificial Intelligence (AI) are used to organize, manipulate, and use the sources of speech knowledge (Barr et. al, 1981).

Several speech understanding systems have been developed. These include earlier systems such as the HEARSAY-I (Reddy, 1973), DRAGON (Baker, 1975), and SPEECHLIS (Woods, 1975). A comparison of these systems can be found in Reddy (1976). More recent systems include the HEARSAY-II (Erman, 1980), HARPY (Lowerre et. al, 1980), and HWIM (Wolf, 1980). A comparison of these latter systems is given by Barr et. al (1981). These systems emphasized different problems in speech understanding research and systems design. Description of these systems is beyond the scope of the present study.
3.3. **Automatic Speech Recognition for the Deaf**

The advances made in the technology of automatic speech recognition have a great potential for the communication needs of the deaf persons. Automatic speech recognition will enable a TDD to recognize the spoken word and translate it into printed characters for the deaf individuals. This will eliminate the use of another TDD by a normal hearing person on the other end of the line (Stoker, 1982).

With successful recognition of the speech of the deaf persons, the oral communication abilities of the profoundly deaf persons could be greatly expanded. Automatic speech recognition with a visual display could serve as a communication aid for the deaf persons (Flanagan, 1982). Such an aid would enable normal hearing individuals to understand deaf speech and deaf individuals to recognize normal speech. This would also enable the deaf individuals to operate voice-controlled machines around home and work. The present work investigates that possibility, that is, to use the techniques of automatic speech recognition to identify and recognize the speech of the deaf. The AIRDS for Automatic Identification and Recognition of Deaf Speech is to be developed.
Automatic speech recognition could also be used in speech training for the deaf persons. Templates of words and sounds are stored in the recognizer for the deaf persons to practice pronouncing. The SIRENE system features the use of automatic speech recognition in speech training for the deaf persons (Haton, 1980).

3.4. Speaker Recognition:

Speaker recognition is a term used to refer to identifying people from their voices, that is, based on their voice characteristics. This is also called speaker identification or verification. The problem of automatic speaker recognition is to classify an unknown utterance as belonging to, or having been spoken by, one of a set of reference speakers. This is a decision making process that uses some features of the speech signal to determine if a particular person is the speaker of that utterance. This problem is divided into two parts; namely: measurements (or feature extraction) and classification (Atal, 1976).

3.4.1. Measurements:

In the first part, a number of measurements are made on the speech signal to provide a set of features representing the speaker-identifying information of
speech. There are two kinds of speaker-identifying features. These include high level features such as dialect, context, and style of speech and low level (acoustic) features such as intensity, pitch, and formants. As yet, the high level features have not been used in speaker recognition systems because of practical difficulties in acquiring, quantifying, and using such information. Instead, the techniques of automatic speaker recognition focus on the low level acoustic features (Doddington, 1985).

Wolf (1972) listed the ideal characteristics for speaker-identifying features. First, they should be efficient, that is, they should show high inter-speaker variability — variations due to anatomical differences such as the size of the vocal tract or due to speaking habits such as the temporal characteristics of speech. Second, they should be stable over time and environment, that is, they should show low intra-speaker variability — variations due to differences in situations such as in speaking rates, emotional states, or environment of speaker. Finally, they should be easy to measure, difficult to mimic, and occurring in speech frequently and naturally.
Speech intensity, pitch, spectrum, predictor coefficients, formants, and timing and speaking rates have been listed as important sources of speaker identifying information and they have been found useful in speaker recognition systems (Atal, 1976). The speech spectrum was used by Furui et. al (1972). Predictor coefficients were used by Sambur (1976) and Atal (1974). Pitch and intensity along with correlation coefficients were used by Markel et. al (1977). Pitch contours were used by Atal (1974), Wolf (1972), and Doddington (1973).

3.4.2. Classification:

Classification is a decision making process to determine if a particular person is the speaker of the given utterance. Using some appropriate decision rule such as the minimum distance classification rule, a distance score is computed from the measurements made on the speech signal, and the utterance is assigned to the speaker with the minimum score. This score represents the amount of dissimilarity between a feature vector (the set of measurements) representing the input utterance and a reference vector of the speaker.

The reference vector of a speaker is the mean vector of a number of vectors representing utterances produced by that speaker. These utterances are called the training
patterns and the process of obtaining the reference vector is called training. Several possible measures such as the Euclidean and Covariance weighting distances are used to compute the distance score in speaker recognition systems (Atal, 1976).

In the present study, the problem of identifying deafness in speech will be considered as a special case of automatic speaker recognition problem in which only two speakers (or actually two classes of speakers: normal and deaf) are recognized. In chapter 4, a set of features that reflect the deviations between normal and deaf speech will be used. Also, a nonlinear discriminant function rule, based on the Euclidean distance, will be used to make the classification decision.

3.5. Linear Predictive Coding of Speech

Linear Predictive Coding (LPC) analysis is one of the most powerful techniques of digital signal processing, and it has become one of the most important tools in automatic speech recognition. LPC is now the predominant method for estimating speech parameters such as spectra, pitch, formants, and vocal tract area function. In addition, LPC can be used in speech segmentation as well as in determination of pattern similarity. These and other concepts of LPC will be discussed next.
3.5.1. Linear Prediction Model:

The basic mathematical idea behind the method of linear predictive coding of speech is that a digital speech sample can be approximated by a linear combination of past values of the signal. Let the predicted sample of $u(n)$ be

$$\hat{u}(n) = - \sum_{i=1}^{M} a(i) u(n-i)$$  \hspace{1cm} 3.1

where $M$ defines the order of the LPC model, and $a(i)$ are the LPC parameters.

The error between the actual sample $u(n)$ and the predicted one $\hat{u}(n)$ is given by

$$e(n) = u(n) - \hat{u}(n)$$

$$= \sum_{i=1}^{M} a(i) u(n-i)$$  \hspace{1cm} 3.2

The parameters $a(i)$ are determined in such a way so that the total squared error summed over a finite time interval is minimized with respect to each of the parameters. The total squared error is given by

$$E = \sum_{n0}^{n1} [e(n)]^2$$
where \( n_0 \) and \( n_1 \) define the finite interval of time. The criterion for defining this time interval is known as the short time principle and is discussed later in this section. Typically, \( n_0 = 0 \) and \( n_1 = N-1 \) where \( N \) is the number of samples in that interval.

Define the correlation matrix \( C_{1,1} \) by

\[
C_{1,1} = \sum_{n_0}^{n_1} u(n-i) u(n-j)
\]

Then,

\[
E = \sum_{i}^{M} \sum_{j}^{M} a(i) C_{1,1} a(j)
\]

Now \( E \) is minimized by setting the partial derivatives of \( E \) with respect to the coefficients \( a(j) \) to zero, that is,

\[
\frac{\partial E}{\partial a(j)} = 0 = 2 \sum_{i}^{M} a(i) C_{1,1}
\]

which are \( M \) simultaneous equations in \( M \) unknowns \( a(i) \), and \( a(0) \) is set to 1. These equations are solved to obtain the desired predictor parameters \( a(i) \) by one of two
methods: the autocorrelation method and the covariance method. These parameters are also called the filter or autoregressive coefficients which model the speech signal as an all-pole filter. This model can also be specified in terms of the autocorrelation coefficients or reflection coefficients which are closely related to the predictor parameters.

3.5.2. **Speech Representation Using LPC:**

Given a set of predictor coefficients, other parameteric representations of speech such as spectra, pitch, formants, and vocal tract area function can be derived. Smooth spectral envelopes are computed from the output of the Fourier transform of the filter coefficients (Markel, 1976; and Rabiner, 1978). The formant frequencies are estimated from the smooth spectral envelope by finding the location of the spectral peaks (Markel, 1973). Pitch period is computed from the autocorrelation function or by inverse filtering of the speech signal (Markel, 1972). Finally, the vocal tract area function is also computed by inverse filtering of the speech signal (Wakita, 1973).

3.5.3. **Analysis Conditions:**

Before using the method of LPC, certain analysis parameters must be set properly. These parameters
include the sampling frequency, number of filter coefficients, number of speech samples in each frame, and pre-emphasis factor. The sampling frequency is dependent on the frequency range of the speech signal and is set to a value higher than twice the maximum frequency in the speech band. A typical value of 10 KHz is used in most speech recognition systems.

The number of filter coefficients is dependent on the number of frequency peaks or resonant frequencies in the speech band where two coefficients are used to represent each peak in addition to two or four coefficients to represent the spectral characteristics of the excitation source. Typical values of 10 to 14 are used in speech recognition systems.

The criterion for setting the frame length is known as the short time principle. The frame length must not be so long as to smear together successive, yet distinctive speech events, but it must be long enough to yield sufficient frequency resolution and to produce smoothly varying function of time. For spectral analysis, more than one pitch period is needed, while more than two pitch periods are required for pitch estimation. Typical values of 200 to 500, representing 20 to 50 milliseconds of speech at 10 KHz sampling rate, are used
in recognition systems. Successive analysis frames may overlap or they may be shifted by small intervals. Typical values of 5 to 12 milliseconds are used (Wolf, 1980).

Pre-emphasizing the speech signal is used to enhance the spectral peaks in the high frequency region and, therefore, recommended for formant analysis. Typically, a fixed first order digital filter with coefficient set from 0.95 to 0.98 is used in speech processing systems (Wakita, 1981).

3.5.4. **LPC in Speech Recognition**

The method of linear predictive coding is very useful for automatic speech recognition. It is especially effective for acoustic feature extraction as described above. Also, certain LPC representation allow for simple distance measures that can be used in determining the pattern similarity (Itakura, 1975). In addition, LPC-based feature vectors have been used in speaker recognition systems (Sambur, 1976; and Atal, 1974). Further, LPC can be used for speech segmentation (Wakita, 1981), and for voiced-unvoiced-silence classification (Atal et. al, 1976).
Atal (1976) listed the reasons for which the use of LPC coefficients in speaker recognition systems is desirable. First, it eliminates the necessity of deciding as to which the speech characteristics would be suitable for recognition. The predictor coefficients represent the combined information about speech characteristics. Second, being independent of pitch and intensity information, an LPC-based recognition systems could be used to evaluate other systems based on pitch and intensity information. Finally, LPC analysis could be easily implemented by digital hardware.
CHAPTER IV

METHODOLOGY

In this chapter, the method used to identify and recognize utterances produced by deaf speakers is described. First, the theory and formulations of this work are presented. Identification of deafness in speech is treated as a special case of an automatic speaker recognition problem in which only two speakers (or actually two classes of speakers) are recognized. Next, the speech of deaf persons, because of its nature, is recognized as strings of connected words by using a two-step hybrid recognizer. Second, implementation of the basic components of the AIRD system; namely: the speech preprocessor, speech classifier, speech recognizer, and trainer using a PDP-11/23 minicomputer with the Interactive Laboratory System (ILS) software is described. Finally, the experimental procedure used to test the AIRD system is presented.

4.1. Theory and Formulations:

The application of the techniques of automatic speech recognition to identify and recognize the speech of deaf persons is not a direct one. A set of measurements must
first be made to represent the characteristics of deaf speech and its deviations from normal speech. This is not an easy task due to the large variability in deaf speech even for the same speaker. In addition, the segmentation method used for recognition of deaf speech must distinguish between word boundaries and long intra-word pauses that are often inserted in deaf speech. The algorithms used to overcome these problems are described next.

4.1.1. **Identification of Deafness in Speech**

4.1.1.1. **Introduction**

To identify deafness in speech is to classify a given speech utterance into one of two classes: normal and deaf speech. This is an automatic speaker recognition problem in which only two speakers (actually two classes of speakers) are recognized based on measurements made on the speech signal. According to the discussion given in sections 2.2 and 3.4, the measurements used in the present study represent four speaker-dependent features. These features are intonation contours, formant transitions, timing and speaking rates, and intensity contours. These features were shown to reflect the characteristics of deaf speech and their deviations from normal speech. They offer sufficient speaker-dependent information that can be used effectively to distinguish between normal and deaf speech.
speech. To review:

1. Intonation is either insufficient or excessive in the speech of the deaf persons. In general, the average value of the fundamental frequency is lower in the speech of deaf speakers than the normal hearing male speakers. The average range of change in the fundamental frequency in deaf speech is about one fourth the change in normal speech in some cases (insufficient intonation) and about twice the change in normal speech in others (excessive intonation).

2. In the speech of deaf persons, transitions of all formants in general and the second in particular are reduced in both time and frequency. The average values and ranges of change in all formants are lower in deaf speech than in normal speech.

3. More and longer inter-, and intra-word pauses are inserted in deaf speech than in normal speech. In addition, deaf speakers take, on the average, two to three times longer than normal hearing speakers do to produce the same utterance.

4. The average value and the range of change of the signal energy are lower in the speech of deaf persons than
in normal speech.

The above four features are used to provide a set of 10 measurements for speech classification. First, the input utterance is divided into quasi-syllable speech segments by marking drops in the signal energy. Then, these measurements, which constitute the elements of the feature vector representing the input utterance, are computed for each segment and will be denoted as:

- $\bar{F}_0_m$ = average value of the fundamental frequency
- $\Delta F_0_m$ = standard deviation of fundamental frequency
- $\bar{F}_1_m$ = average value of first formant
- $\Delta F_1_m$ = standard deviation of first formant
- $\bar{F}_2_m$ = average value of second formant
- $\Delta F_2_m$ = standard deviation of second formant
- $\bar{F}_3_m$ = average value of third formant
- $\Delta F_3_m$ = standard deviation of third formant
- $\overline{SEG}_m$ = average speech segment duration
- $\overline{SIL}_m$ = average silence segment duration

The present study also investigates the use of an LPC-based feature set to identify deafness in speech. In this case, the feature vector consists of 11 predictor coefficients from which a 10th order LPC model is derived. The process of obtaining this vector is described in section 3.5. In either case, a distance score is
computed from the feature vector and the utterance is assigned to the class with the minimum distance. A nonlinear discriminant function rule, based on the Euclidian distance, is used to make the classification decision. This is described next.

4.1.1.2. Speech Classification:

To show how the classification decision is made, let the feature vector representing the unknown utterance be

\[ X = [x(1), x(2), \ldots, x(N)]^T \]

where \( N \) is the number of measurements. Also, let \( A \) and \( B \) represent the two possible pattern classes where \( A \) is for normal speech and \( B \) is for deaf speech. Now define

\[ D_A(X) \] as the discriminant function associated with class \( A \),

and

\[ D_B(X) \] as the discriminant function associated with class \( B \)

such that

if the input utterance represented by the feature vector \( X \) is in class \( A \), that is, \( X \) is produced by a normal hearing speaker, then the value of \( D_A(X) \) is less than the value of \( D_B(X) \); and
if the input utterance is in class B, that is, produced by a deaf speaker, then the value of \( D_b(X) \) is less than the value of \( D_a(X) \). Mathematically, this can be rewritten as

\[
\text{if } X \in A \quad \text{then} \quad D_a(X) < D_b(X)
\]

and

\[
\text{if } X \in B \quad \text{then} \quad D_b(X) < D_a(X)
\]

Now define \( D_a(X) \) and \( D_b(X) \) as nonlinear combinations of the measurements \( x(i) \); that is,

\[
D_a(X) = \sum_{i=1}^{N} [x(i) - a(i)]^2 \quad 4.2
\]

and

\[
D_b(X) = \sum_{i=1}^{N} [x(i) - b(i)]^2 \quad 4.3
\]

The values of \( a(i) \) and \( b(i) \); \( i=1,2, \ldots, N \); are characteristics of the domains representing the two classes A and B respectively. Each domain is defined by utterances produced by speakers of the same class. These values constitute the reference vectors \( R_a \) and \( R_b \) representing these two classes and given by

\[
R_a = [a(1), a(2), \ldots, a(N)]^T \quad 4.4.a
\]

\[
R_b = [b(1), b(2), \ldots, b(N)]^T \quad 4.4.b
\]
The decision boundary between class A and class B is in the form

\[ D(X) = D_A(X) - D_B(X) \]  

4.5

The classification decision is determined by the value of \( D(X) \) such that

if \( D(X) < 0 \) then \( X \in A \)  

4.6.a

and

if \( D(X) > 0 \) then \( X \in B \)  

4.6.b

This nonlinear discriminant function rule described above does not depend on the form of distribution of the measurements. For this reason, and since there are only two classes to recognize, this rule is used in the present study.

4.1.1.3. Generation of the Reference Vectors:

This classification decision, that is, correct identification of the speaker of the input utterance, is possible only if the values of the reference vectors \( a(i) \) and \( b(i) \); \( i=1,2, \ldots, N \); are known. The process of training or learning is used to obtain the best estimates of these values by observing patterns of input speech utterances with known classifications. These utterances,
referred to as the training patterns, will also be used to adjust the values of the reference vectors so that the system adapts to new words or new speakers.

To show how the reference vectors are obtained, let $I_A$ and $I_B$ be two sets of training patterns belonging to the two possible speech classes $A$ and $B$ respectively. Each element in these sets is defined by a feature vector $X$ as given by equation 4.1.

For correct classification, there exists two reference vectors $R_A$ and $R_B$ such that

$$ D(X) < 0 \quad \text{for all} \quad X \in I_A $$

and

$$ D(X) > 0 \quad \text{for all} \quad X \in I_B $$

Now, in the case of wrong classification, that is,

- if $D(X) > 0$ for any $X \in I_A$
  
  then $R_A = R_A + qX$ \hspace{1cm} 4.7.a

- and

- if $D(X) < 0$ for any $X \in I_B$
  
  then $R_B = R_B - qX$ \hspace{1cm} 4.7.b
This procedure is repeated until correct classification is achieved. Before training, the constant q, referred to as the correction increment, is set to any fixed positive value while a(i) and b(i) are set to any convenient values.

4.1.2. Recognition of the Speech of Deaf Persons

4.1.2.1. Introduction

The speech of the deaf persons is, as described in section 2.2, characterized by the insertion of frequent and long inter-, and intra-word pauses. As a result, an utterance produced by a deaf speaker would appear like a string of connected words produced by a normal hearing speaker (where articulations at word boundaries are limited). This would suggest the use of the techniques of connected word recognition to recognize the speech of deaf persons. However, one must be careful not to detect the intra-word pauses in deaf speech as word boundaries resulting in wrong recognition. The recognizer must distinguish between word boundaries and inserted intra-word pauses to avoid inaccurate segmentation.

In the present study, a two-step word-segmentation approach is employed to recognize deaf speech. In the first step, and before recognition, a number of candidate word boundaries are determined by detecting drop gaps in
the signal energy. These could be true word boundaries, (inter-word pauses), or just gaps caused by intra-word pauses in deaf speech. As a result, the input speech utterance is divided into a string of speech segments separated by pauses. These segments could be whole words or parts of words.

In the second step, the true boundaries are determined during recognition. Each speech segment as well as all possible groupings of consecutive speech segments are matched against all reference patterns. The word grouping which yields the the best match determines the recognition results. This implicitly determines the best choice of word boundaries so that an integer number of nonoverlapping words fit between the beginning and the ending of the utterance. A string match algorithm is used to find the best word grouping.

4.1.2.2. The String Match Algorithm

Assume that there is a speech utterance produced by a deaf speaker where a number of intra-, and inter-word pauses are inserted resulting in, say I, speech segments separated by I-1 candidate boundaries. Also assume that these segments can be grouped into a number of words, say J, where each word is a concatenation of a number of consecutive speech segments, say K. The value of K varies
from 1 up to some limit determined by the maximum number of syllables in a vocabulary word.

For example, consider the case in Figure 4 in which we have a deaf utterance divided into three speech segments (i.e., \( i = 3 \)). One possible word grouping results in three words (i.e., \( j = 3 \)) of one segment each. Two other possible grouping result in two words (i.e., \( j = 2 \)) in each — one grouping with one segment in the first word and two segments in the second word, and another grouping with two segments in the first word and one segment in the second word. The last possible grouping results in only one word (i.e., \( j = 1 \)) of three segments. However, if the number of segments in a word does not exceed two, then this last grouping will be excluded.

In general, assume that for a particular grouping \( g \), \( g = 1, 2, \ldots, G \); there are \( J \) words labelled as

\[
W(1), W(2), \ldots, W(J)
\]

where \( G \) is the number of all possible groupings (\( G \) is 4 in the above example). For each word, a word match distance is computed by comparing its pattern with all reference patterns of the vocabulary words. Let these reference patterns be labelled as
where $Q$ is the number of words in the vocabulary. The word match distance of $W(j)$ is given by

$$D(j) = \min_{q} D(W(j), R(q))$$  \hspace{1cm} 4.9$$

where $D(W(j), R(q))$ represents some distance measure between the $j$th word and the $q$th reference pattern and will be discussed later in this section. The index of the reference pattern that best matches the word $W(j)$ is given by

$$q^*(j) = \arg\min_{q} D(W(j), R(q))$$  \hspace{1cm} 4.10$$

A grouping score is now computed as the sum of $J$ word match distances as follows

$$DIS(g) = \sum_{j} D(j)$$  \hspace{1cm} 4.11$$

The best word grouping is chosen as the one with the minimum score as

$$DIS = \min_{g} DIS(g)$$  \hspace{1cm} 4.12$$

and the index of that grouping is given by
\[ g^* = \underset{g}{\text{argmin}} \text{DIS}(g) \]  

As a result, the integer number of words that best fits between the beginning and end of the utterance is given by \( J \), and recognition of each word is given by equation 4.10.

The minimization of equation 4.12 is accomplished recursively by dynamic programming through a string match algorithm. A partial string score is computed as

\[ \text{PDIS}(j) = \min_{K} [ \text{PDIS}(j-1) + D(j) K(j) ] \]  

where \( j \) is the index of the \( j \)th word in the utterance string, and \( D(j) \) is its word match distance given by equation 4.9. This word is a concatenation of \( K(j) \) speech segments, where \( K(j) \) varies from 1 and up to some limit \( K_{\text{max}} \) which is defined by the maximum number of syllables in a word in the vocabulary (\( K_{\text{max}} \) is 4 in the present system). The concatenation that yields the minimum score defines the \( j \)th word, its boundary, and the partial string score through it.

\( \text{PDIS}(0) \) is the initial partial score at the beginning of the utterance, and is set to zero. \( K(j) \) is also used as a weighting function to normalize all words or partial strings to the same basis, that is, the number of
syllables in that word.

Then, for a given number of words, \( J \), the optimum string score is given by the partial score at the end of the utterance, that is, at \( j=J \)

\[
\text{DIS} = \text{PDIS}(J) \quad 4.15
\]

4.1.2.3. The Word Match Distance:

The word match distance \( D(j) \) represents the best similarity measure between the candidate word \( W(j) \) and all reference patterns \( R(q) \) as given by equation 4.9 which is repeated here as

\[
D(j) = \min_{q} D(W(j), R(q))
\]

where \( D(W(j), R(q)) \), written as \( D(j,q) \) for simplicity, is the distance measure between the \( j^{th} \) word and the \( q^{th} \) reference pattern. This distance is computed by using the method of dynamic time warping (DTW) to compensate for differences in speaking rates.

First, assume that the word \( W(j) \) consists of a sequence of \( M \) vectors representing \( M \) time frames, that is,

\[
W(j) = \{ w_j(1), w_j(2), \ldots, w_j(M) \} \quad 4.16
\]
where \( w_j(m) \) is the feature representation of the \( m^{th} \) frame in \( W(j) \). In the present system, this is a vector of 11 predictor coefficients from which a \( 10^{th} \) order LPC model is derived. The process of obtaining this vector from the speech signal is described in section 3.4.

Similarly, assume that the reference template consists of a sequence of \( N \) vectors, that is,

\[
R(q) = \{ r_q(1), r_q(2), \ldots, r_q(N) \}
\]

Now define the local distance between the \( m^{th} \) frame in \( W(j) \) and the \( n^{th} \) frame in \( R(q) \) as

\[
d(m,n) = \| w_j(m) - r_q(n) \|_2.
\]

This distance is minimum when the two frames corresponds to each other. A time warping path (or alignment function) is used to define the frames in the reference pattern \( R(q) \) for which correspondence is expected. The total word distance \( D(j,q) \) is then computed as the sum of the minimum local distances for all frames in the word \( W(j) \) and is given by

\[
D(j,q) = \min_N \sum_M d(m,n)
\]
where \( n = p(m) \) is the warping path (alignment function) that maps frames in \( W(j) \) onto the corresponding frames in \( R(q) \). The optimal warping path defines the distance \( D(j,q) \) which can be computed recursively by dynamic programming.

Define an accumulated distance function \( D_\alpha(m,n) \) as the total distance from the start point (the first frame in the word, \( m=1 \) and \( n=1 \)) to a point \( m,n \) along the path, that is,

\[
D_\alpha(m,n) = D_\alpha(m-1,n-1) + \min \left\{ d(m-1,n) \cdot g(m,n), \right. \\
\left. d(m-1,n-1), d(m-1,n-2) \right\}  
\]

where \( g(m,n) \) is a nonlinear weighting function used to guarantee that the optimum path does not stay flat for two consecutive frames and is given by

\[
g(m,n) = 1 \ \text{if} \ p(m) = p(m-1) \\
\text{and} \\
g(m,n) = \infty \ \text{if} \ p(m) \neq p(m-1) 
\]

The word distance is obtained at the end of the word, that is, at \( m=M \) and \( n=N \) and is given by

\[
D(j,q) = D_\alpha(M,N) 
\]
4.1.2.4. Generation of the Reference Patterns

In the present system, the reference patterns $R(q)$ ($q=1, 2, ..., Q$) are generated by using the method of dynamic time warping to combine or average several tokens of a word to form a reference pattern or a template of that word. These tokens are repetitions of that word separated by long silence pauses.

Assume that there is $V$ tokens for the $q^{th}$ word in the vocabulary. First, the reference pattern $R(q)$ is chosen as the token closest to the average duration of all tokens of that word. Next, this reference pattern is time-warped to all other tokens resulting in $V-1$ warped tokens. This procedure is repeated until only one warped token is obtained which represents the composite reference template of that word.

4.2. The AIRDS System

The AIRDS system for automatic identification and recognition of deaf speech is shown in the block diagram in Figure 5. It consists of four functional blocks; namely: the Preprocessor, the Classifier, the Recognizer, and the Trainer.

In the first block, speech features are extracted from the raw speech signal of the input utterance to produce a speaker-identifying feature vector for speech
Figure 4. Possible word groupings in a 3-segment utterance.
Figure 5. Blockdiagram of the AIRDS system (S is the input speech signal, and D is the output display).
classification and a feature pattern for speech recognition. In the second block, the speaker identifying feature vector is used to identify deafness in speech; that is, to classify the input utterance into normal or deaf speech.

In the third block, the feature pattern of the input utterance is used to recognize the speech message contained in that utterance. A flag coming from the Classifier is used by the Recognizer to determine which set of reference patterns (normal or deaf) is used for recognition.

In the fourth block, these reference patterns are generated by the Trainer and stored in the system memory also shown in the diagram. The trainer is also used to generate the reference vectors by which the Classifier can adapt to new speakers or to new utterances. Implementation of these four blocks using a PDP-11/23 minicomputer with the Interactive Laboratory System (ILS) software is described next.

4.2.1. Speech Preprocessor:

Speech Preprocessor is that part of the AIRDS system by which speech features are extracted from the raw speech signal of the input utterance using the method of linear
predictive coding of speech. These features form the input to the other parts of the system; namely: the Classifier, Recognizer, and Trainer. These features are the predictor coefficients, fundamental frequency, formant frequencies, and energy.

A flowchart of the Preprocessor is shown in Figure 6. First, the speech signal is digitized at 10 K Hz sampling rate using the REC command of ILS (refer to Appendix A for description of ILS). This command performs analog to digital conversion on the speech signal of the input utterance into a file that is previously specified by the FIL command. The file name, file length (recording time), and sampling rate are specified by the FIL command and accessed by the REC command from the file header.

Recording time is specified by the number of frames and the number of samples per frame (referred to as the context). The context is set by the CTX command or by the FIL command. Normal single channel operation is used for recording, and each sample is stored in a 16-bit computer word. The recording process will be considered again in section 4.4.

The second step in the Preprocessor is to initialize and properly set the analysis parameters required for LPC analysis. These parameters are the pre-emphasis factor
Figure 6. Flowchart of the Preprocessor.
(PR), the number of points in the analysis window (N), the type of windowing function (HM), the number of spectral peaks (NP), and the number of predictor coefficients (M). Based on the discussion given in section 3.5, and using the INA command of ILS, these parameters are preset to the following values:

- PR = 98
- N = 256
- HM = Hamming Window is used
- NP = 3
- M = 10

The last step in the Preprocessor is to extract the speech features from the stored speech samples of the input utterance. The following procedure is used:

1. The API command for pitch estimation using LPC analysis is executed. The resulting analysis file has a number of analysis vectors equal to the number of frames in the input utterance file. The analysis vector consists of 128 locations (refer to the Appendix for description of the analysis vector). The predictor coefficients, fundamental frequency, and signal energy are computed and stored in locations 1 to 11, 119, and 123; respectively as a result of executing the API command. A voicing flag is also stored in location 115, while a flag
The speech-silence classification is stored in location 113.

2. The RSO command is then executed to compute the formant frequencies where the number of formants is specified by the number of spectral peaks (NP). As a result, the three formant frequencies are stored in locations 31, 34, 37; the bandwidths in locations 32, 35, 38; and the amplitudes in locations 33, 36, 39; respectively.

3. The SRE command is then used to obtain the feature vectors required for speech classification. One feature vector is formed from 11 predictor coefficients, and is obtained from locations 1 to 11 in the analysis vector. This vector will be referred to as the linear predictive feature vector (LPV). Another feature vector is obtained from locations 123, 119, 31, 34, and 37 which contain the signal energy, fundamental frequency, and the three formants; respectively. This second vector will be referred to as the acoustic feature vector (AFV).

4. The QUR command, which is designed to extract features from labeled speech segments, is used to obtain a feature pattern of the input utterance to be used for speech recognition. This pattern is stored as a matrix.
with the number of vectors equal to the number of frames in the input utterance, and each vector consists of 11 predictor coefficients. This pattern, which is a compact representation of the LPC feature vector obtained above, can also be used for speech classification.

4.2.2. Speech Classifier:

Speech Classifier is that part of the AIRDS system by which an input utterance is identified as deaf or normal speech. Speech classification, as described in section 4.1.1, consists of comparison of a feature vector representing the input utterance with two reference vectors representing the two speech classes. A distance score is computed for each class, and the utterance is assigned to the class with the minimum distance. This feature vector is provided by the Preprocessor as either the AFV or LPV vectors described above.

A flowchart of the Classifier is shown in Figure 7. First, the SME command is used to obtain the measurements proposed in section 4.1 for speech classification from the AFV vector resulting in a modified AFV vector of 10 elements. This vector is denoted as \( X \) in the flowchart. The BPA command is next used to compute the discriminant functions \( D_A(X) \) and \( D_B(X) \) associated with the two speech classes using the Euclidean distance as given by equations
Figure 7. Flowchart of the Classifier.
4.2 and 4.3.

The reference vectors \( R_a \) and \( R_m \) are obtained by the Trainer to be described later. Then, the decision boundary given by equation 4.5 is used to assign the utterance to one of the two classes. \( X \) may also represent the LPV vector for speech classification using predictor coefficients. In the flowchart, \( F \) is a flag used to indicate if the input utterance has been classified as normal or deaf speech. This flag is used later by the Recognizer and the Trainer to identify the speech class of the input utterance.

4.2.3. **Speech Recognizer**

The Recognizer is that part of the AIRDS system by which the message contained in the classified utterance is recognized. This is a hybrid recognizer which employs a two-step word segmentation method as described in section 4.1.2. In the first step, a number of candidate word boundaries are determined by detecting drop gaps in the signal energy. These could be true boundaries or just gaps due to intra-word pauses. As a result, the input utterance is divided into a string of speech segments that could be whole words or parts of words. In the second step, a string match algorithm is used to compare each speech segment as well as all possible groupings of
cosecutive speech segments against all reference templates. The word grouping which yields the best match determines the recognition results and implicitly the word boundaries.

A flowchart of the Recognizer is shown in Figure 8. First, the DSP command is used to locate the silent frames, and hence the beginning and ending frames of each speech segment. This can be also done automatically by testing the signal energy (location 123 in the analysis vector) against a prespecified threshold, or by tracking the speech-silence flag stored in location 113.

These speech segments are then labelled by using the LBF and LBA commands. The LBF is first used to create a label file for the input utterance, that is, to contain the labels given to the speech segments in that utterance. These labels are specified by the LBA command. The label of a segment consists of six fields which contain information about that segment. Only the first three fields are respectively used for speech class, segment identity, and segment number.

Segment grouping (or concatenation) is done by using the TRE command which can be used to merge two or more input record files into one output file containing the
Figure 8. Flowchart of the recognition algorithm.
feature representation of a candidate word. The BPA command is then used to compute the word match distance for each concatenation by matching it with the reference patterns P. The flag F, which is stored in FIELD-1 of the label of each segment, is used to determine which set of patterns is used for matching. This distance is then used to compute the partial string score given by equation 4.14. The concatenation which gives the minimum score determines the recognition results for that word as well as its boundary which is defined by the last segment in this concatenation. The use of the TRE and BPA commands is repeated to recognize the next word in the string until the end of the utterance is reached. This is obtained when the label of the last segment in the recognized word matches the label given to the last segment in the utterance.

4.2.4. The Trainer:

The Trainer is that part of the AIRDS system by which the Classifier and the Recognizer can adapt new speakers or new words. Training the Classifier is done by observing the training patterns (patterns of input utterances with known classifications) to adjust the values of the reference vectors RA and RB as described in section 4.1.3. A flowchart of this procedure is shown in Figure 9.
Figure 9. Flowchart of training the Classifier.
In this flowchart, \( X \) is the feature representation of a training pattern. At the beginning of the training procedure, the reference vectors \( R_a \) and \( R_b \) are set to any convenient values, and the correction increment \( q \) is set to any fixed positive value. The BPA command is used to compute the decision boundary \( D(X) \) as given by equations 4.2, 4.3, and 4.5. The flag \( F \) is determined by the speech class of the input utterance which is stored in FIELD-1 of its label.

Training the Recognizer is done by observing several tokens of a word, and then combining them to form or modify the reference pattern \( R(q) \) of that word as described in section 4.2.4. A flowchart of this procedure is shown in Figure 10. The duration of the token is determined by the length (in number of frames) of the file containing it. The WRP command is used to time warp each token to the selected reference pattern. The flag \( F \), which is stored in FIELD-1 of the label of the input utterance, is used to determine which class this pattern belongs to.

4.3. Experimental Method:

4.3.1. Subjects:

The subjects used in the present study were two deaf male speakers referred to as DS1 and DS2, and two normal
Figure 10. Flowchart of training the Recognizer.
hearing male speakers referred to as NS1 and NS2. The selection criteria for the deaf speakers were: (1) the subjects have a congenital bilateral hearing loss, (2) comparable oral speech training background to minimize intra-speaker variability, (3) subjects received oral training from an early age, and (4) ability of subjects to articulate the speech samples adequately to insure above average intelligibility. Details of age, etiology, training, and hearing loss of the two deaf subjects appear in Table 1.

The two normal hearing speakers, referred to as NS1 and NS2, were used as the control subjects in this study. Both subjects were males and they were 22 and 27 years old respectively. These two speakers did not have any evidence of speech problem or hearing loss. In addition, their speech is a representative of a single dialect area (the Midwest area).

4.3.2. Speech Material:

The speech samples used in the present study were obtained from a subset of an artificial recognition language previously designed and used for airline information and reservation systems. A list of these utterances, which represent 31 isolated words and 8 connected-word strings, is given in Table 2. The
| AGE (YEARS) | 27 | 35 |
| GENDER | MALE | MALE |
| ETIOLOGY OF HEARING LOSS | CONGENITAL | CONGENITAL |
| PURE-TONE AVERAGE (dB) | | |
| AT 500 Hz | 85 | >100 |
| AT 100 | 100 | |
| AT 2000 | 110 | |
| SPEECH TRAINING STARTED AT AGE | 18 MONTH | 15 MONTH |
| PERIOD OF SPEECH TRAINING | 6 YEARS | 10 YEARS |
| INTELLIGIBILITY SCORE* | 21.36% | 17.94% |

*Intelligibility score for each speaker was computed by two inexperienced listeners as the ratio of the number of recognized words to the total number of words.
TABLE 2. Speech materials.

1. ISOLATED WORDS:

<table>
<thead>
<tr>
<th>ONE-SYLLABLE WORDS</th>
<th>2-SYLLABLE WORDS</th>
<th>3-SYLLABLE WORDS</th>
<th>4-SYLLABLE WORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1. in</td>
<td>W25. return</td>
<td>W28. Columbus</td>
<td>W29. reservation</td>
</tr>
<tr>
<td>W2. the</td>
<td>W26. morning</td>
<td></td>
<td>W30. information</td>
</tr>
<tr>
<td>W3. make</td>
<td>W27. nonstop</td>
<td></td>
<td>W31. Cincinnati</td>
</tr>
<tr>
<td>W4. a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W5. I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W6. want</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W7. some</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W8. please</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W9. need</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W10. would</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W11. like</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W12. first</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W13. class</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W14. flight</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W15. seat</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W16. will</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W17. pay</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W18. cash</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W19. how</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W20. much</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W21. is</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W22. fare</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W23. from</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W24. to</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. CONNECTED-WORDS STRINGS:

P1. I need some information please.
P2. I want a nonstop flight.
P3. From Columbus to Cincinnati.
P4. I would like a first class seat.
P5. I want to make a reservation.
P6. I want to return in the morning.
P7. How much is the fare?
P8. I will pay in cash.
isolated words include 24 mono-syllable words, 3 two-syllable words, one three-syllable word, and 3 four-syllable words. The connected-word strings include one phrase, 6 sentences, and one question.

These samples were selected for two reasons. First, they cover a wide range of the language; from short mono-syllable words to long sentences, and from a declarative statement to a question. They provide linguistically real situations which reflect the changes in speech characteristics. Second, they represent a practical application of a speech communication system for airline information and reservation tasks over a telephone line.

4.3.3. Recording Procedure:

All recordings were made in an anechoic chamber with the recording system shown in Figure 11. The speech signal was first recorded on a JVC KD-6 stereo cassette tape recorder and then replayed at the same recording speed where it was next amplified by a Harmon/Kardon PM 640 amplifier, band-pass filtered by a 2KH Krohn-Hite model 3750 filter, and digitized using a KV-11 analog-to-digital converter. The digitized samples were then stored in files on a RL02 disk.
Figure 11. A block diagram of the recording system.
The REC command of ILS was used to perform the analog-to-digital conversion of the speech signal into a file which was previously specified by the FIL command. Normal single channel operation was used for recording, and the speech signal was stored as full 12-bit samples, one per 16-bit computer word.

Before making a recording, each subject was instructed to read the list of words and phrases, shown in Table E, to ensure their familiarity with each item on the list. The microphone was placed about 10 inches from the subject mouth, and the subject was instructed to produce the utterance as naturally as possible and at a moderate intensity level. The subject was then asked to read each item three times with pauses of about 10 seconds between any two trails and the recordings were made. A total of 117 utterances (three times 31 words and 8 phrases) were produced by each subject.

4.3.4. Analysis Procedure:

File specification by the FIL command includes the file name, file length (number of frames), and context (number of points per frame). File name was specified as WDSNNI where

WD denotes ILS data files,
NN denotes utterance number,
I denotes the token number, and S denotes the subject number.

The parameter NN was given the values from 0 to 31 for isolated words and 91 to 98 for connected-word strings.

File length was specified as the number of frames needed to contain the input utterance in addition to few silence frames at the beginning and end of the utterance. This simulates the process of endpoint detection which was done interactively by using the DSP command. The context which also defines the file length was set to 128 points per frame. As a result, 10 frames were used to contain an utterance of about 100 milliseconds long.

In ILS notations, these files were considered as primary files which contained the sampled data. For each file, a secondary analysis file was created to contain the analysis vectors resulting from executing the API and RSO commands. A secondary record file was also created to contain the feature representation of each utterance as described in section 4.2.1. These record files are then considered as primary record files in later steps of the analysis procedure, and other secondary record files are created. The analysis procedure involves file
Figure 12. Flowchart of the analysis procedure.
manipulation, statistical analysis, and distance computation. A flowchart of the analysis procedure is shown in Figure 12.

File manipulation is done by using the TRE, MRE, and MDF commands. The TRE command allows transferring records between files. It also allows merging of record files which can be used in concatenating pattern representations of consecutive speech segments. The MRE command is used to transfer individual records or elements of records between files. The MDF command is used to change values of specific elements in the record file.

Statistical analysis is performed by using the SME, HIS, and PCD commands. The SME command allows computations of means and standard deviations, correlation and covariance, eigen values and vectors, maximum likelihood parameters, in addition to other statistical parameters. The HIS command is used to compute and display the discrete probability distribution (histogram) of selected elements in the record file. It also compares this histogram with a known distribution using the Chi-square goodness of fit. The PCD command is used to perform the principle components analysis to reduce the dimensionality of the data while still retaining most of the variance.
Distance computation for speech classification and recognition is done by using the BPA command as described before. The VDI command is used to compute the inter- and intra-class distances which provide information on correct acceptance and rejection respectively.

Other commands such as the LRE and PLR are used to display the results. The LRE command prints the contents of a record file on the screen or line printer. The PLR command plots a two dimensional display of the data in a record file. It also performs a two dimensional factor analysis of the data and displays the plot in the form of major and minor axes of an ellipse sized at two standard deviations.

4.4. Computation of the Vocal Tract Area Function:

The linear prediction model as described in section 3.5.1. was shown to be equivalent to the acoustic tube model. In this model, the vocal tract is represented as a concatenation of a finite number of cylindrical sections of equal lengths. The cross-sectional areas are computed from a set of predictor coefficients which are computed directly from the acoustic speech signal (Wakita, 1979).
The vocal tract area function (VTAF), that is the area of the vocal tract as a function of the distance from the glottis, is used in the present work to study the articulatory gestures of deaf speakers and their differences from those of normal hearing speakers. This will help establish a set of articulatory parameters that could be used to distinguish between deaf and normal speech.

Based on the discussion given in section 2.3, the speaker's articulatory gestures at two positions are studied. These are: the tongue position (back of the tongue) and the larynx position (above the glottis). While the first is used to study the posture characteristics of speech production, the second is used to study the interarticulatory coordination during speech production.

The VTR command of ILS is used to compute 11 cross-sectional areas (derived from a 10th order LPC model) representing the vocal tract from the lips to the glottis. The area at the glottis is taken as a reference and is normalized to 1 square millimeters. This command will be also used to plot mid-sagittal representations based on the values of the area function.
In the present study, an attempt will be made to study the changes in the tongue and larynx positions from one frame to another for normal and deaf speech. A set of measurements will be used to quantify these changes in the two speech classes. These measurements will next be used to identify deafness in speech as described in section 4.1.
CHAPTER V

RESULTS AND DISCUSSIONS

This chapter is divided into four sections. In the first section, phonatory results are presented. In the second section, results of speech classification to identify deafness in speech are given. Recognition results are presented in the third section. Results on the use of the vocal tract area function for classification are presented in the last section.

5.1. Phonatory Results :

5.1.1. Comparison of normal and deaf speech samples :

Displays of the utterances W11 (one-syllable word), W27 (two-syllable word), and P7 (connected-word sentence) are shown in Figures 13, 15, and 17 respectively for a normal hearing speaker, and in Figures 14, 16, and 18 respectively for a deaf speaker. These utterances are "like", "nonstop", and "How much is the fare?"; respectively.
Figure 13. Acoustic features of the utterance W1 produced by a normal hearing speaker.
Figure 14. Acoustic features of the utterance W27 produced by a normal hearing speaker.
Figure 15. Acoustic features of the utterance P7 produced by a normal hearing speaker.
Figure 16. Acoustic features of the utterance 'W11' produced by a deaf speaker.
Figure 17. Acoustic features of the utterance W27 produced by a deaf speaker.
Figure 18. Acoustic features of the utterance P7 produced by a deaf speaker.
In these figures, the speech signal is displayed at the top, the fundamental frequency (dotted) and energy (solid) in the middle, and the three formant frequencies at the bottom. The tic marks on the time scale represent 10 millisecond intervals. The tic marks on the vertical scale at the bottom represent 500 Hz intervals. The tic marks on the vertical axis in the middle represent 50 Hz intervals for the fundamental frequency.

**Intonation**

Normal and deaf subjects demonstrated relatively different patterns of intonation contours. All patterns, however, are characterized by a sudden rise at the beginning of each speech segment (syllable) and a sudden decline at its end. Within each speech syllable, the deaf subject showed slightly smaller changes in the fundamental frequency than that for the normal hearing speaker. In addition, the deaf speaker demonstrated smaller changes in the average fundamental frequency from one speech segment to another. This is shown in Figure 19 for the utterance P7. These results are summarized in Table 3. The reader must be reminded here that these values were computed for quasi-syllable speech segments as described in chapter 4.
Table 3. Fundamental frequency values (Hz).

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th></th>
<th>Deaf</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F_{0m}$</td>
<td>$F_{0\alpha}$</td>
<td>$F_{0m}$</td>
<td>$F_{0\beta}$</td>
</tr>
<tr>
<td>W11</td>
<td>148.7</td>
<td>45.5</td>
<td>106.3</td>
<td>58.1</td>
</tr>
<tr>
<td>W27</td>
<td>111.3</td>
<td>59.4</td>
<td>117.0</td>
<td>52.0</td>
</tr>
<tr>
<td></td>
<td>125.5</td>
<td>54.6</td>
<td>110.5</td>
<td>45.8</td>
</tr>
<tr>
<td>P7</td>
<td>161.9</td>
<td>17.6</td>
<td>107.8</td>
<td>58.1</td>
</tr>
<tr>
<td></td>
<td>143.2</td>
<td>63.5</td>
<td>127.5</td>
<td>18.8</td>
</tr>
<tr>
<td></td>
<td>26.3</td>
<td>32.4</td>
<td>106.2</td>
<td>48.3</td>
</tr>
<tr>
<td></td>
<td>77.9</td>
<td>64.9</td>
<td>24.5</td>
<td>50.4</td>
</tr>
<tr>
<td></td>
<td>97.0</td>
<td>54.3</td>
<td>120.0</td>
<td>40.4</td>
</tr>
</tbody>
</table>
Figure 19. Intonation contours for the utterance P7.
In general, these results agree with the findings of previous studies (Osberger et al., 1982). However, the ratio of the change in deaf speech to that in normal speech is lower in the present study when compared to those obtained in previous studies.

Formants

Formant transitions represented a great deviation between normal and deaf speech in the speech samples used in the present study. While normal hearing subjects demonstrated transitions across speech segments, deaf speakers showed transitions both across (in some cases) and within the speech segments. Gradual changes of the formant frequencies within speech segments were demonstrated particularly by the normal hearing speakers.

The average values of the formant frequencies are lower in deaf speech than in normal speech. Transitions in the formant frequencies are also lower in deaf speech than in normal speech as expected. This is demonstrated in Figure 20 for the utterance P7. These results are summarized in Table 4.
These results are also in general agreement with previous findings (Monsen, 1979). However, the transitions within the speech segments in deaf speech are surprising. It can only be said that it did not occur in producing all utterances, and that it might have been caused by inability of the deaf speaker to articulate those speech segments.

**Timing and speaking rates**

Comparing Figures 13 with 14, 15 with 16, and 17 with 18, the deaf speaker took 2.36, 2.42, and 1.76 times longer than the normal hearing speaker did in producing the utterances W11, W27, and P7; respectively. This shows that the speaking rate of the normal hearing speaker is about two times faster than the deaf speaker. Examining these figures also shows that the deaf speaker demonstrated a prolongation of both speech and silence segments when compared to the normal hearing speaker. These results, which are shown in Table 5, are also in agreement with previous findings (Osberger et. al, 1982; and Nickerson, 1975).
Table 4. Formant frequencies (Hz).

<table>
<thead>
<tr>
<th></th>
<th>Normal 1</th>
<th></th>
<th>Normal 2</th>
<th></th>
<th>Normal 3</th>
<th></th>
<th>Deaf 1</th>
<th></th>
<th>Deaf 2</th>
<th></th>
<th>Deaf 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(F_{1m})</td>
<td>(F_{2m})</td>
<td>(F_{3m})</td>
<td>(F_{1m})</td>
<td>(F_{2m})</td>
<td>(F_{3m})</td>
<td>(F_{1m})</td>
<td>(F_{2m})</td>
<td>(F_{3m})</td>
<td>(F_{1m})</td>
<td>(F_{2m})</td>
<td>(F_{3m})</td>
</tr>
<tr>
<td>W11</td>
<td>836.3</td>
<td>2032.7</td>
<td>3438.5</td>
<td>588.9</td>
<td>2290.4</td>
<td>3750.3</td>
<td>588.9</td>
<td>2290.4</td>
<td>3750.3</td>
<td>588.9</td>
<td>2290.4</td>
<td>3750.3</td>
</tr>
<tr>
<td>W27</td>
<td>801.8</td>
<td>2084.0</td>
<td>3733.0</td>
<td>519.4</td>
<td>2366.0</td>
<td>3994.6</td>
<td>519.4</td>
<td>2366.0</td>
<td>3994.6</td>
<td>519.4</td>
<td>2366.0</td>
<td>3994.6</td>
</tr>
<tr>
<td></td>
<td>820.4</td>
<td>2020.2</td>
<td>3444.3</td>
<td>560.2</td>
<td>2100.3</td>
<td>3455.2</td>
<td>560.2</td>
<td>2100.3</td>
<td>3455.2</td>
<td>560.2</td>
<td>2100.3</td>
<td>3455.2</td>
</tr>
<tr>
<td>P7</td>
<td>667.4</td>
<td>1141.4</td>
<td>2768.2</td>
<td>956.1</td>
<td>2131.9</td>
<td>3459.9</td>
<td>956.1</td>
<td>2131.9</td>
<td>3459.9</td>
<td>956.1</td>
<td>2131.9</td>
<td>3459.9</td>
</tr>
<tr>
<td></td>
<td>492.3</td>
<td>1666.1</td>
<td>3014.3</td>
<td>350.6</td>
<td>1458.1</td>
<td>2880.3</td>
<td>350.6</td>
<td>1458.1</td>
<td>2880.3</td>
<td>350.6</td>
<td>1458.1</td>
<td>2880.3</td>
</tr>
<tr>
<td></td>
<td>2267.1</td>
<td>3041.9</td>
<td>3258.9</td>
<td>366.5</td>
<td>2317.9</td>
<td>3464.8</td>
<td>366.5</td>
<td>2317.9</td>
<td>3464.8</td>
<td>366.5</td>
<td>2317.9</td>
<td>3464.8</td>
</tr>
<tr>
<td></td>
<td>712.4</td>
<td>2080.1</td>
<td>2719.6</td>
<td>1014.4</td>
<td>2255.9</td>
<td>3239.4</td>
<td>1014.4</td>
<td>2255.9</td>
<td>3239.4</td>
<td>1014.4</td>
<td>2255.9</td>
<td>3239.4</td>
</tr>
<tr>
<td></td>
<td>544.8</td>
<td>1589.8</td>
<td>2750.8</td>
<td>542.9</td>
<td>1859.1</td>
<td>3238.6</td>
<td>542.9</td>
<td>1859.1</td>
<td>3238.6</td>
<td>542.9</td>
<td>1859.1</td>
<td>3238.6</td>
</tr>
<tr>
<td></td>
<td>(F_{1d})</td>
<td>(F_{2d})</td>
<td>(F_{3d})</td>
<td>(F_{1d})</td>
<td>(F_{2d})</td>
<td>(F_{3d})</td>
<td>(F_{1d})</td>
<td>(F_{2d})</td>
<td>(F_{3d})</td>
<td>(F_{1d})</td>
<td>(F_{2d})</td>
<td>(F_{3d})</td>
</tr>
<tr>
<td>W11</td>
<td>702.9</td>
<td>649.4</td>
<td>746.4</td>
<td>419.5</td>
<td>593.7</td>
<td>771.1</td>
<td>419.5</td>
<td>593.7</td>
<td>771.1</td>
<td>419.5</td>
<td>593.7</td>
<td>771.1</td>
</tr>
<tr>
<td>W27</td>
<td>747.0</td>
<td>760.5</td>
<td>672.6</td>
<td>413.1</td>
<td>510.3</td>
<td>593.2</td>
<td>413.1</td>
<td>510.3</td>
<td>593.2</td>
<td>413.1</td>
<td>510.3</td>
<td>593.2</td>
</tr>
<tr>
<td></td>
<td>720.3</td>
<td>690.4</td>
<td>680.2</td>
<td>420.4</td>
<td>535.3</td>
<td>620.3</td>
<td>420.4</td>
<td>535.3</td>
<td>620.3</td>
<td>420.4</td>
<td>535.3</td>
<td>620.3</td>
</tr>
<tr>
<td>P7</td>
<td>139.7</td>
<td>492.0</td>
<td>193.0</td>
<td>586.6</td>
<td>724.1</td>
<td>690.6</td>
<td>586.6</td>
<td>724.1</td>
<td>690.6</td>
<td>586.6</td>
<td>724.1</td>
<td>690.6</td>
</tr>
<tr>
<td></td>
<td>211.9</td>
<td>591.9</td>
<td>539.6</td>
<td>219.8</td>
<td>451.8</td>
<td>532.0</td>
<td>219.8</td>
<td>451.8</td>
<td>532.0</td>
<td>219.8</td>
<td>451.8</td>
<td>532.0</td>
</tr>
<tr>
<td></td>
<td>736.3</td>
<td>471.1</td>
<td>1432.3</td>
<td>323.8</td>
<td>474.7</td>
<td>571.4</td>
<td>323.8</td>
<td>474.7</td>
<td>571.4</td>
<td>323.8</td>
<td>474.7</td>
<td>571.4</td>
</tr>
<tr>
<td></td>
<td>676.8</td>
<td>743.3</td>
<td>1021.7</td>
<td>696.2</td>
<td>387.0</td>
<td>546.3</td>
<td>696.2</td>
<td>387.0</td>
<td>546.3</td>
<td>696.2</td>
<td>387.0</td>
<td>546.3</td>
</tr>
<tr>
<td></td>
<td>445.0</td>
<td>781.3</td>
<td>862.7</td>
<td>350.1</td>
<td>390.8</td>
<td>737.6</td>
<td>350.1</td>
<td>390.8</td>
<td>737.6</td>
<td>350.1</td>
<td>390.8</td>
<td>737.6</td>
</tr>
</tbody>
</table>
Figure 20. Transitions of the second formant for the utterance P7 (N for normal speech and D for deaf speech).
Table 5. Energy levels (volt²) and timing rates (ms).

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Deaf</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\bar{E}_m$</td>
<td>$\bar{E}_{sd}$</td>
</tr>
<tr>
<td>W11</td>
<td>2789.4</td>
<td>1082.9</td>
</tr>
<tr>
<td>W27</td>
<td>2227.8</td>
<td>852.5</td>
</tr>
<tr>
<td></td>
<td>1830.4</td>
<td>533.6</td>
</tr>
<tr>
<td>P7</td>
<td>7618.1</td>
<td>2136.8</td>
</tr>
<tr>
<td></td>
<td>4233.1</td>
<td>1699.5</td>
</tr>
<tr>
<td></td>
<td>2180.0</td>
<td>440.2</td>
</tr>
<tr>
<td></td>
<td>2149.2</td>
<td>1530.0</td>
</tr>
<tr>
<td></td>
<td>1741.4</td>
<td>691.8</td>
</tr>
</tbody>
</table>

$E$ = energy level  
SYdur = average syllable duration
Intensity

Examining these figures also shows that the deaf speaker demonstrated lower levels and ranges of changes in speech intensity as compared to the normal hearing speaker in producing the same utterances. These values are 1427.5 and 474.6 for the deaf speaker as compared to 2537.4 and 584.1 for the normal hearing speaker. This is in general agreement with previous findings.

Again, the reader must be reminded that the values shown in this section were obtained for quasi-syllable segments which were determined as described in chapter 4 by detecting drops in the signal energy. This would account for the surprising small value of F0m which is shown in Figure 19.
5.1.2. Statistical analysis:

Statistical analysis was performed on the acoustic features of the speech signal, namely: the fundamental frequency, formant frequencies, and intensity in addition to the LPC parameters. For each speech class, 31 utterances (W1, W2, ..., W31) were used to represent its sample population. A summary of the mean and standard deviation values is given in Table 6, while the correlation matrices are given in Table 7 for normal speech and Table 8 for deaf speech.

These results show that the average values of the fundamental frequency, formant frequencies, and signal energy are lower in deaf speech when compared to normal speech (t test; p < 0.10). This agrees with the previous studies referred to in chapter 2. The results also show lower variations of the signal energy and first formant frequency in deaf speech than in normal speech. However, they show higher variations in the fundamental frequency and the frequency of the second formant. This is not surprising and it can be accounted for by the much larger variability in deaf speech than in normal speech. Within each utterance, however, the changes are lower for deaf speech when compared to normal speech as described above.
Table 6. Acoustic measurements.

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Deaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F_0_m )</td>
<td>113.70</td>
<td>104.64</td>
</tr>
<tr>
<td>( F_0_ad )</td>
<td>17.06</td>
<td>38.65</td>
</tr>
<tr>
<td>( F_1_m )</td>
<td>691.48</td>
<td>487.23</td>
</tr>
<tr>
<td>( F_1_ad )</td>
<td>309.27</td>
<td>220.37</td>
</tr>
<tr>
<td>( F_2_m )</td>
<td>2094.50</td>
<td>1952.60</td>
</tr>
<tr>
<td>( F_2_ad )</td>
<td>348.79</td>
<td>577.62</td>
</tr>
<tr>
<td>( F_3_m )</td>
<td>3669.20</td>
<td>3385.20</td>
</tr>
<tr>
<td>( F_3_ad )</td>
<td>267.15</td>
<td>985.72</td>
</tr>
<tr>
<td>( E_m )</td>
<td>2537.40</td>
<td>1427.50</td>
</tr>
<tr>
<td>( E_ad )</td>
<td>584.14</td>
<td>474.64</td>
</tr>
<tr>
<td>( SY_{dur} )</td>
<td>200.00</td>
<td>410.00</td>
</tr>
<tr>
<td>( SI_{dur} )</td>
<td>85.00</td>
<td>270.00</td>
</tr>
<tr>
<td>( R1_m )</td>
<td>-0.489</td>
<td>-0.452</td>
</tr>
<tr>
<td>( R1_ad )</td>
<td>0.206</td>
<td>0.198</td>
</tr>
<tr>
<td>( R2_m )</td>
<td>0.252</td>
<td>0.040</td>
</tr>
<tr>
<td>( R2_ad )</td>
<td>0.164</td>
<td>0.271</td>
</tr>
</tbody>
</table>
Table 7. Correlation matrix for normal speech.

<table>
<thead>
<tr>
<th></th>
<th>E</th>
<th>F0</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>R1</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F0</td>
<td>0.23</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>-0.05</td>
<td>0.19</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>0.14</td>
<td>0.03</td>
<td>0.42</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F3</td>
<td>-0.09</td>
<td>0.08</td>
<td>0.25</td>
<td>0.41</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>-0.30</td>
<td>-0.18</td>
<td>0.22</td>
<td>0.22</td>
<td>-0.04</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>0.02</td>
<td>0.16</td>
<td>0.45</td>
<td>0.04</td>
<td>-0.46</td>
<td>0.48</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Table 8. Correlation matrix for deaf speech.

<table>
<thead>
<tr>
<th></th>
<th>E</th>
<th>F0</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>R1</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F0</td>
<td>0.52</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>0.09</td>
<td>0.34</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>-0.05</td>
<td>0.12</td>
<td>0.56</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F3</td>
<td>-0.09</td>
<td>0.07</td>
<td>0.52</td>
<td>0.92</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>-0.24</td>
<td>-0.18</td>
<td>0.19</td>
<td>-0.27</td>
<td>-0.32</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>0.43</td>
<td>0.41</td>
<td>0.47</td>
<td>-0.05</td>
<td>-0.14</td>
<td>0.57</td>
<td>1.0</td>
</tr>
</tbody>
</table>
For deaf speech, the results in Table 8 show correlation values of 0.52 between the energy and formant frequency, 0.38 between the fundamental frequency and the first formant, 0.56 between the first and second formants and 0.52 between the first and third formants, and 0.92 between the second and third formants. They also show smaller correlation of values from 0.1 to 0.5 between these parameters and the LPC parameters. A high correlation of 0.87 was found between the first and second reflection coefficients. Smaller correlation values are shown in Table 7 for normal speech.

Histograms

The discrete probability distributions of the fundamental frequency are displayed in Figure 21 for normal speech and Figure 22 for deaf speech. A statistical summary is printed at the bottom of each display. These figures show that while a normal distribution might be a good fit for the distribution of the fundamental frequency in normal speech, it is not so for deaf speech. The distribution of the fundamental frequency in deaf speech showed no mode as compared to a mode of 111.6 Hz for normal speech. The Chi-square test for goodness of fit resulted in a value of 10.05 for deaf speech as compared to 2.13 for normal speech.
Figure 21. Histogram of the fundamental frequency for normal speech.
Figure 22. Histogram of the fundamental frequency for deaf speech.
Histograms of the second formant frequency and the first reflection coefficient are shown in Figures 23 and 25 for normal speech and in Figures 24 and 26 for deaf speech; respectively. The figures show that although the distributions of these parameters can be approximated by gaussian curves for normal speech as well as for deaf speech, the fit is closer in the case of normal speech as compared to deaf speech. This is demonstrated by the values of the chi-square test that are printed at the bottom of each display.

Principle Components

A two dimensional scatter plot of the second formant frequency (x-axis) vs the first formant frequency (y-axis) is shown in Figure 27, and of the first reflection coefficient (x-axis) vs the second reflection coefficients (y-axis) is shown in Figure 28. In these figures, N stands for normal speech utterances while D stands for deaf speech utterances. Statistical information for each symbol is also displayed. This includes the means and covariance along the x-axis (XMEAN, RXX) and along the y-axis (YMEAN, RYY), the covariance (RXY), the normalized cross correlation (RHO), and the eigen values and slopes of the first and second principle components (EIGEN1, SLOPE1, EIGEN2, and SLOPE2).
Figure 23. Histogram of the second formant frequency for normal speech.
Figure 24. Histogram of the second formant frequency for deaf speech.
Figure 25. Histogram of the first reflection coefficient for normal speech.
Figure 26. Histogram of the first reflection coefficient for deaf speech.
Figure 27. 2D scatter plot of the first and second formant frequencies.
Figure 28. 2D scatter plot of the first and second reflection coefficients.
These plots show that although the two speech classes overlap in the space defined by the first and second formants or in the space defined by the first and second reflection coefficients, speech classification is still possible. Two dimensional scatter plots of the fundamental frequency vs the second formant, and of the fundamental frequency vs the first reflection coefficient also show the same results. These plots are shown in Figures 29 and 30 respectively.

Scatter plots obtained for two different sets of utterances one produced by the normal hearing speakers and the other produced by the deaf speakers are shown in Figures 31 to 34. The first set consists of the words W1 to W15, while the second set consists of the words W16 to W30. These figures show that the classification task is easier in this case than in the case of classifying the same set of utterances in the two speech classes.
Figure 29. 2D scatter plot of the fundamental frequency and second formant frequency.
Figure 30. 2D scatter plot of the fundamental frequency and first reflection coefficient.
Figure 31. 2D scatter plot of the first and second formant frequencies using different sets of utterances for normal and deaf speech.
Figure 32. 2D scatter plot of the first and second reflection coefficients using different sets of utterances for normal and deaf speech.
Figure 33. 2D scatter plot of the fundamental frequency and second formant frequency using different sets of utterances for normal and deaf speech.
Figure 34. 2D scatter plot of the fundamental frequency and first reflection coefficient using different sets of utterances for normal and deaf speech.
5.2. Classification Results:

5.2.1. Classification of isolated words:

For each speaker, 93 utterances were classified first by using the acoustic feature vector (AFV) representation and second by using the LPC feature vector (LPV) representation. These utterances were also used to generate the reference vectors as described in chapter 4.

The results are summarized in Tables 9 and 10 respectively. They show that comparable classification rates were obtained for both cases. These are 87.10% for AFV and to 83.87% for LPV. The results also show that while higher rates were obtained in classifying normal speech using LPV, similar rates were obtained for both classes using AFV.

Using AFV, each of the utterances W8, W18, W21, and W31 were misclassified more than one time as deaf speech when produced by normal hearing speakers, while the utterances W2, W3, W4, and W6 were misclassified more than one time each as normal when produced by deaf speakers.

Using LPV, each of the utterances W1, W7, W21, and W31 were misclassified more than one time as deaf speech when produced by normal hearing speaker, while the
Table 9. Classification results using AFV.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Normal</th>
<th>Deaf</th>
<th>Accuracy</th>
<th>Accuracy%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS1</td>
<td>84</td>
<td>9</td>
<td>90.32</td>
<td>87.10</td>
</tr>
<tr>
<td>NS2</td>
<td>78</td>
<td>15</td>
<td>83.87</td>
<td>87.10</td>
</tr>
<tr>
<td>DS1</td>
<td>6</td>
<td>87</td>
<td>93.55</td>
<td>87.10</td>
</tr>
<tr>
<td>DS2</td>
<td>18</td>
<td>75</td>
<td>80.64</td>
<td>87.10</td>
</tr>
</tbody>
</table>
Table 10. Classification results using LPV.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Normal</th>
<th>Deaf</th>
<th>Accuracy%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS1</td>
<td>82</td>
<td>11</td>
<td>88.17</td>
</tr>
<tr>
<td>NS2</td>
<td>80</td>
<td>13</td>
<td>86.02</td>
</tr>
<tr>
<td>DS1</td>
<td>15</td>
<td>78</td>
<td>83.87</td>
</tr>
<tr>
<td>DS2</td>
<td>21</td>
<td>72</td>
<td>77.42</td>
</tr>
</tbody>
</table>
utterances W6, W13, W14, W18, W19, and W22 were misclassified more than one time each as normal speech when produced by deaf speakers.

In both cases, the utterance W6 was misclassified when produced by deaf speakers, while the utterances W21 and W31 were misclassified when produced by normal hearing speakers.

5.2.2. Classification of connected-word strings:

For each speech class, 48 utterances were classified first by using the acoustic feature vector (AFV) representation and then by using the linear prediction feature vector (LPV) representation. These utterances are up to 2.56 seconds long for normal speech and up to 6.0 seconds long for deaf speech. For each utterance, successive longer segments were classified to study the effect of the utterance length on the classification results. The results are shown in Figure 35 for normal speech and Figure 36 for deaf speech.

These figures demonstrate that the classification rate for normal speech as well as for deaf speech increased successively as longer parts of the utterance were considered. They also show that higher classification accuracies were obtained for normal speech
as compared to deaf speech. For normal speech, utterance durations of about 0.5 second long were required to achieve a maximum accuracy of about 88%. For deaf speech, however, durations of about 0.9 seconds long resulted in a maximum classification rate of about 84%. This is demonstrated by the larger slopes in figure 35 as compared to the slopes in figure 36.

The figures also demonstrate that comparable results were obtained by using both the AFV and LPV representations. The classification rates are 88% and 86% respectively for normal speech, and 84% and 83% respectively for deaf speech. However, the classification task using the LPV representation was about two to three times faster than the case when the AFV representation was used. This is explained by the fact that the acoustic features were extracted from the LPC parameters.

5.2.3. Classification results for new utterances:

The performance of the system was next evaluated by using different sets of utterances for training (odd number words: W1, W3, ... ) and testing (even number words: W2, W4, ... ). The classification rates were 81.11% and 78.88% for normal and deaf speech respectively when AFV was used, and 81.11% and 77.77% for normal and
deaf speech respectively when LPV was used. These rates are lower than those obtained in section 5.2.1. using the same set of utterances for training and testing.

5.2.4. **Classification results for new speakers**

The utterances W11, W27, and W5 were produced three times each by an independent speaker (a third deaf speaker whose utterances were not used for training the system). These were next used as keywords to test the system performance on an independent data. With one exception, they were correctly classified as deaf speech when AFV and LPV were used. The utterance W5 was misclassified once as normal speech when LPV was used.

These results do not mean that higher rates are obtained for classifying utterances produced by independent deaf speakers. However, they show that the system could be used to identify deafness at least in keywords produced by these deaf speakers.
Figure 3b. The effect of utterance length on classification rate for normal speech (x using AFV and o using LPV).

Figure 3b. The effect of utterance length on classification rate for deaf speech (x using AFV and o using LPV).
5.3. Recognition Results:

5.3.1. Isolated-word recognition:

For each speech class, a total of 186 utterances (31 words each produced 3 times by each speaker) were used to test the recognizer in the AIRDS system. For each word, all 6 tokens were used to generate the reference template of that word as described in chapter 4. Confusion matrices are shown in Table 11 for normal speech and Table 12 for deaf speech. The results demonstrated higher recognition rate of 97.85% for normal speech as compared to a rate of 93.01% for deaf speech.

For normal speech, the results in Table 11 show that the following confusions occurred:

- W1 was recognized as W21 once.
- W4 was recognized as W17 once.
- W23 was recognized as W7 two times.

For deaf speech, the results in Table 12 show that the following confusions occurred:

- W4 was recognized as W17 once.
- W7 was recognized as W23 once.
- W9 was recognized as W15 two times.
- W10 was recognized as W16 once.
- W14 was recognized as W11 once.
- W17 was recognized as W4 once.
Table 11. Isolated-word recognition (normal speech).

<table>
<thead>
<tr>
<th>Word</th>
<th>Recognized As</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>5</td>
</tr>
<tr>
<td>W2</td>
<td>6</td>
</tr>
<tr>
<td>W3</td>
<td>5</td>
</tr>
<tr>
<td>W4</td>
<td>6</td>
</tr>
<tr>
<td>W5</td>
<td>6</td>
</tr>
<tr>
<td>W6</td>
<td>6</td>
</tr>
<tr>
<td>W7</td>
<td>6</td>
</tr>
<tr>
<td>W8</td>
<td>6</td>
</tr>
<tr>
<td>W9</td>
<td>6</td>
</tr>
<tr>
<td>W10</td>
<td>6</td>
</tr>
<tr>
<td>W11</td>
<td>6</td>
</tr>
<tr>
<td>W12</td>
<td>6</td>
</tr>
<tr>
<td>W13</td>
<td>6</td>
</tr>
<tr>
<td>W14</td>
<td>6</td>
</tr>
<tr>
<td>W15</td>
<td>6</td>
</tr>
<tr>
<td>W16</td>
<td>6</td>
</tr>
<tr>
<td>W17</td>
<td>6</td>
</tr>
<tr>
<td>W18</td>
<td>6</td>
</tr>
<tr>
<td>W19</td>
<td>6</td>
</tr>
<tr>
<td>W20</td>
<td>6</td>
</tr>
<tr>
<td>W21</td>
<td>6</td>
</tr>
<tr>
<td>W22</td>
<td>2</td>
</tr>
<tr>
<td>W23</td>
<td>6</td>
</tr>
<tr>
<td>W24</td>
<td>6</td>
</tr>
<tr>
<td>W25</td>
<td>6</td>
</tr>
<tr>
<td>W26</td>
<td>6</td>
</tr>
<tr>
<td>W27</td>
<td>6</td>
</tr>
<tr>
<td>W28</td>
<td>6</td>
</tr>
<tr>
<td>W29</td>
<td>6</td>
</tr>
<tr>
<td>W30</td>
<td>6</td>
</tr>
<tr>
<td>W31</td>
<td>6</td>
</tr>
</tbody>
</table>

Total 97.85
Table 12. Isolated-word recognition (deaf speech).

| W1 | 6 | 100 |
| W2 | 6 | 100 |
| W3 | 6 | 100 |
| W4 | 6 | 83.3 |
| W5 | 6 | 100 |
| W6 | 6 | 100 |
| W7 | 5 | 100 |
| W8 | 1 | 83.3 |
| W9 | 5 | 100 |
| W10 | 5 | 100 |
| W11 | 5 | 100 |
| W12 | 5 | 100 |
| W13 | 5 | 100 |
| W14 | 5 | 100 |
| W15 | 5 | 100 |
| W16 | 5 | 100 |
| W17 | 5 | 100 |
| W18 | 5 | 100 |
| W19 | 5 | 100 |
| W20 | 5 | 100 |
| W21 | 2 | 66.7 |
| W22 | 1 | 66.7 |
| W23 | 1 | 66.7 |
| W24 | 1 | 66.7 |
| W25 | 1 | 66.7 |
| W26 | 1 | 66.7 |
| W27 | 1 | 66.7 |
| W28 | 1 | 66.7 |
| W29 | 1 | 66.7 |
| W30 | 1 | 66.7 |
| W31 | 1 | 66.7 |

total 93.01
- W21 was recognized as W1 two times.
- W22 was recognized as W17 once.
- W23 was recognized as W7 once.
- W29 was recognized as W30 once.
- W30 was recognized as W29 once.

These results show that no consistent patterns of errors were noted in recognizing either speech class. It can be seen, however, that there are three groups of words where confusions occurred in both speech classes. These are W1 with W21, W4 with W17, and W7 with W23. For deaf speech, confusions in other word groups such as W29 and W30 were also noted. These results only emphasize the fact that the variability in deaf speech is larger than that in normal speech resulting in more confusions and lower recognition rates in deaf speech as compared to normal speech.

5.3.2. Connected-word recognition:

For each speech class, a total of 48 utterances (8 phrases each produced three times by each speaker) were used in this experiment. Recognition accuracies for each class are summarized in Table 13. These accuracies were computed as the ratio of the correctly recognized words to the total number of words. The overall recognition accuracies are 85.61% for normal speech and 81.81% for
Confusion matrices of these results are shown in Table 14 for normal speech and Table 15 for deaf speech. It is seen from Table 14 that the following confusions occurred most frequently in normal speech:

- W1 was recognized as W21 3 times out of 12 tries.
- W4 was recognized as W17 4 times out of 18 tries.
- W23 was recognized as W7 2 times out of 6 tries.

These confusions are consistent with the results obtained in section 5.3.1 for isolated-word recognition. This shows that these confusions occurred due to the similarities that exist between the words W1 and W21, W4 and W17, and W23 and W7.

Other confusions such as in recognizing the words W5, W6, and W24 occurred due to the difficulty in segmenting these words especially when they are combined together, as in the phrases P2, P5, and P6. These confusions are:

- W5 was recognized as either W1, W2, W3, W17, W21, or W22 9 times out of 36 tries.
- W6 was recognized as either W10, W16, or W3 5 times out of 18 tries.
- W24 was recognized as either W4, W10, W16, or W19 7 times out of 24 tries.
Table 13. Summary of connected-word recognition.

Recollection results %

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>words</td>
<td>errors</td>
<td>rate</td>
<td></td>
<td>errors</td>
<td>rate</td>
</tr>
<tr>
<td>P1</td>
<td>30</td>
<td>5</td>
<td>83.33</td>
<td></td>
<td>6</td>
<td>80.00</td>
</tr>
<tr>
<td>P2</td>
<td>30</td>
<td>3</td>
<td>90.00</td>
<td></td>
<td>4</td>
<td>86.66</td>
</tr>
<tr>
<td>P3</td>
<td>24</td>
<td>3</td>
<td>87.50</td>
<td></td>
<td>3</td>
<td>87.50</td>
</tr>
<tr>
<td>P4</td>
<td>42</td>
<td>8</td>
<td>80.95</td>
<td></td>
<td>9</td>
<td>78.57</td>
</tr>
<tr>
<td>P5</td>
<td>36</td>
<td>6</td>
<td>83.33</td>
<td></td>
<td>8</td>
<td>77.77</td>
</tr>
<tr>
<td>P6</td>
<td>42</td>
<td>8</td>
<td>80.95</td>
<td></td>
<td>9</td>
<td>78.57</td>
</tr>
<tr>
<td>P7</td>
<td>30</td>
<td>3</td>
<td>90.00</td>
<td></td>
<td>5</td>
<td>83.33</td>
</tr>
<tr>
<td>P8</td>
<td>30</td>
<td>2</td>
<td>93.33</td>
<td></td>
<td>4</td>
<td>86.66</td>
</tr>
<tr>
<td>TOTAL</td>
<td>264</td>
<td>38</td>
<td>85.61</td>
<td></td>
<td>48</td>
<td>81.81</td>
</tr>
</tbody>
</table>
Table 14. Connected-word recognition (normal speech).

<table>
<thead>
<tr>
<th>Recognized as</th>
<th>T</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>W1 9</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>W2 11</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>W3 5</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>W4 14</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>W5 1 2 1 21</td>
<td>36</td>
<td>9</td>
</tr>
<tr>
<td>W6 1 5 2 2 2 1</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>W7 6</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>W8</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>W9</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>W10</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>W11</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>W12</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>W13</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>W14</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>W15</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>W16</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>W17 1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>W18</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>W19</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>W20</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>W21</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>W22</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>W23</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>W24 2 2 1 1 3 4 11</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>W25</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>W26</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>W27</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>W28</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>W29</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>W30</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>W31</td>
<td>6</td>
<td>0</td>
</tr>
</tbody>
</table>

Total 264 38
Table 15. Connected-word recognition (deaf speech).

<table>
<thead>
<tr>
<th>Recognized as</th>
<th>T E R</th>
<th>R R</th>
<th>R E</th>
<th>B R</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W2</td>
<td>10</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W3</td>
<td></td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W5</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>W6</td>
<td>1</td>
<td>13</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>W7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W10</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W21</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>264</td>
</tr>
</tbody>
</table>

48
The other 8 confusions that occurred in connected-word recognition of normal speech are probably caused due to speech variability. These accounted for only 3.03% of the errors in this experiment as compared to 3.41% of the errors caused by word similarity and 7.95% of the errors due to segmentation difficulty.

It is also seen from Table 15 that the following confusions occurred most frequently in deaf speech:

- W1 was recognized as W21 3 times out of 18 tries.
- W4 was recognized as W17 5 times out of 18 tries.
- W7 was recognized as W23 once out of 6 tries.
- W17 was recognized as W4 once out of 6 tries.
- W21 was recognized as W17 2 times out of 6 tries.
- W23 was recognized as W7 2 times out of 6 tries.

These confusions are also consistent with the results obtained for isolated-word recognition. They occur probably due to the word similarities which exist between the words W1 and W21, W4 and W17, and W7 and W23. They account for 5.30% of the errors in connected word recognition of deaf speech. This is higher than the corresponding value of 3.41% obtained for normal speech. This shows that normal speakers produce similar words more efficiently than deaf speakers do.
Other confusions in deaf speech occurred in recognizing the words W5, W6, and W24 due to segmentation difficulties. These confusions are:

- W5 was recognized as either W1, W2, W3, W17, W21, or W22 10 times out of 36 tries.
- W6 was recognized as either W10, W16, or W3 5 times out of 18 tries.
- W24 was recognized as either W4, W16, or W19 5 times out of 18 tries.

These confusions accounted for 7.57% of the errors in connected word recognition of deaf speech. This is lower than the corresponding value of 7.95% obtained for normal speech. This shows that segmentation of deaf speech is relatively easier than that of normal speech. This is expected due to the fact that deaf speech contains longer pauses between words than in normal speech.

The other 14 errors that occurred in connected-word recognition of deaf speech are probably caused due to speech variability. This accounted for 5.30% of the errors occurred in deaf speech as compared to 3.03% of the errors occurred in normal speech. This shows that the variability in deaf speech is larger than it is in normal speech.
To summarize, the errors in connected-speech recognition occurred due to:

<table>
<thead>
<tr>
<th></th>
<th>normal</th>
<th>deaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>word similarity</td>
<td>3.41%</td>
<td>5.30%</td>
</tr>
<tr>
<td>segmentation difficulty</td>
<td>7.95%</td>
<td>7.57%</td>
</tr>
<tr>
<td>speech variability</td>
<td>3.03%</td>
<td>5.30%</td>
</tr>
</tbody>
</table>

5.3.3. Evaluation of the 2-step recognizer

The performance of the two-step segmentation method was evaluated by comparing the results obtained in section 5.3.1 and 5.3.2 with recognition results obtained by using a one-step segmentation method (these two methods were discussed in chapter 3). The results of this evaluation is summarized in Table 16. They demonstrate that while the same recognition accuracy were obtained using the two methods for mono-syllable words, higher accuracies were obtained by using the 2-step method for multi-syllable words and connected-word utterances. The results also show that the recognition accuracy decreases as the number of the syllables in the utterance increases.
Table 16. Evaluation of the two-step recognizer.

<table>
<thead>
<tr>
<th></th>
<th>1-syllable words</th>
<th>2-syllable words</th>
<th>3-syllable words</th>
<th>4-syllable words</th>
<th>5-syllable phrases</th>
<th>6-syllable phrases</th>
<th>7-syllable phrases</th>
<th>AVERAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>normal</td>
<td>deaf</td>
<td>normal</td>
<td>deaf</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>one-step</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-syllable</td>
<td>97.22</td>
<td>92.36</td>
<td>97.22</td>
<td>92.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-syllable</td>
<td>94.44</td>
<td>90.00</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-syllable</td>
<td>88.15</td>
<td>83.33</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-syllable</td>
<td>83.33</td>
<td>83.33</td>
<td>93.75</td>
<td>88.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-syllable</td>
<td>77.77</td>
<td>77.77</td>
<td>90.00</td>
<td>84.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phrases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6-syllable</td>
<td>77.77</td>
<td>71.07</td>
<td>83.33</td>
<td>77.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phrases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7-syllable</td>
<td>71.07</td>
<td>69.69</td>
<td>80.95</td>
<td>75.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phrases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AVERAGE</td>
<td>84.25</td>
<td>81.07</td>
<td>92.18</td>
<td>88.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.4. The Vocal Tract Area Function (VTAF)

5.4.1. Introduction

The linear prediction model as described in section 3.5.1 was shown to be equivalent to the acoustic tube model. In this model, the vocal tract is represented as a concatenation of a finite number of cylindrical sections of equal lengths. The cross-sectional areas are computed from a set of predictor coefficients which are computed directly from the acoustic speech signal (Wakita, 1979).

The vocal tract area function (VTAF), that is the area of the vocal tract as a function of the distance from the glottis, is used in the present work to study the articulatory movements of deaf speakers and their differences from those of normal hearing speakers. This will help establish a set of articulatory parameters that can be used to distinguish between deaf and normal speech.

VTR command of ILS is used to compute 11 cross-sectional areas (derived from a 10th order LPC model) representing the vocal tract from the lips to the glottis. The area at the glottis is taken as a reference and is normalized to 1 square millimeters.
Based on the discussion in section 2.3, the speaker's articulatory movements at two positions on the vocal tract are studied. These were: the tongue position (back of the tongue) and the larynx position (above the glottis). While the first was used to study the posture characteristics of speech production, the second was used to study the inter-articulatory coordination during speech production.

5.4.2. Articulatory movements in deaf speakers:

Displays of the vocal tract area at the tongue position for the utterances W11, W27, and P7 are shown in Figures 37, 38, and 39; respectively. In these figures, the x-axis represents the time in milliseconds while the y-axis represents the vocal tract area in square millimeters (the area at the glottis is normalized to 1 square millimeters).

By examining these figures, the following observations are made:

1. For the normal hearing speaker, the number of peaks of the tongue movement is equal to the number of syllables in the utterance: one in W11, two in W27, and 5 in P7. However, this is not so for the deaf speaker where the number of peaks in the tongue movement is always greater than the number of syllables.
Figure 37. The vocal tract area at the tongue position for the utterance /i:/ (a) normal speech and (b) deaf speech.
Figure 38. The vocal tract area at the tongue position for the utterance W27: (a) normal speech and (b) deaf speech.
Figure 39. The vocal tract area at the tongue position for the utterance P7: (a) normal speech and (b) deaf speech.
2. The deaf speaker demonstrated smaller ranges of the tongue movement in producing the utterances W11 and W27 than the normal hearing speaker did. However, the range of the tongue movement in producing the utterance P7 was larger for the deaf speaker as compared to that of the normal hearing speaker.

3. The deaf speaker demonstrated longer durations of the tongue movement in producing all utterances than the normal hearing speaker did.

4. There is no other consistent pattern of the tongue movement that distinguishes the two speakers in producing these utterances.

Displays of the vocal tract area at the larynx position for the utterances W11, W27, and P7 are shown in Figures 40, 41, and 42; respectively. Examining these figures demonstrate the following:

1. The deaf speaker showed more vibration (fast motions) in the larynx movement in producing all utterances than the normal hearing speaker did.

2. The deaf speaker showed smaller ranges of the larynx movement in producing W11 and W27 and larger range in producing P7 than those of the normal hearing speaker.
Figure 40. The vocal tract area at the larynx position for the utterance W11: (a) normal speech and (b) deaf speech.
Figure 41. The vocal tract area at the larynx position for the utterance W27: (a) normal speech and (b) deaf speech.
Figure 42. The vocal tract area at the larynx position for the utterance P7: (a) normal speech and (b) deaf speech.
3. The deaf speaker demonstrated longer durations of the larynx movement in producing all utterances than the normal hearing speaker did.

Examining the vocal tract area at the tongue and the larynx positions for other utterances confirmed the above results. These results demonstrate that the most dramatic articulatory differences observed between the normal hearing and deaf subjects are the duration and range of movement of the tongue and the larynx articulators. These differences are more pronounced in mono-syllable utterances (especially vowels such as "I" and "a") than in multi-syllable utterances (especially connected speech). These results are summarized in Table 17.

5.4.3. The use of VTAF for classification:

The attempt made here is to identify deafness in speech by using a small number of articulatory measurements estimated from the speech signal. Based on the results obtained in section 5.4.2, these measurements are: the duration and range of movement of the tongue and larynx to represent the position and shape of these articulatory organs in the vocal tract.
Using this simple articulatory model, the mono-syllable utterances (W1, W2, ..., W24) were classified. Classification rates of 75.0% and 70.83% were obtained for normal and deaf speech respectively. These results demonstrate the greater variability in deaf speech than in normal speech.

5.4.4. Acoustic-articulatory correspondence:

Correlation values between the formant frequencies and the V1AF were computed to study the acoustic-articulatory correspondence in the two speech classes. The results are shown in Table 18 for normal speech and in Table 19 for deaf speech.

In general, these results suggest that the acoustic-articulatory correspondence in the two speech classes is not a simple one. While the first and second formants may be associated with the throat (A4 to A6), the third formant is associated with the mouth (A3). The larynx (A8 to A9) has an important role in forming the three formants. The glottis (A11), on the other hand, has a minimal role in their formation.
Table 17. Summary of articulatory movements.

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th></th>
<th>Deaf</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>L</td>
<td>D</td>
<td>T</td>
</tr>
<tr>
<td>W11</td>
<td>16.2</td>
<td>2.6</td>
<td>250</td>
<td>10.5</td>
</tr>
<tr>
<td>W27</td>
<td>29.5</td>
<td>2.8</td>
<td>600</td>
<td>16.7</td>
</tr>
<tr>
<td>P7</td>
<td>24.5</td>
<td>1.2</td>
<td>2550</td>
<td>30.2</td>
</tr>
<tr>
<td>Average</td>
<td>23.4</td>
<td>2.2</td>
<td></td>
<td>19.1</td>
</tr>
</tbody>
</table>

- T: Tongue movement (mm$^2$)
- L: Larynx movement (mm$^2$)
- D: Utterance duration (m sec)
Table 18. Acoustic-articulatory correlation for normal speech.

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1 (LIPS)</td>
<td>0.219</td>
<td>0.219</td>
<td>-0.043</td>
</tr>
<tr>
<td>A2</td>
<td>0.453</td>
<td>0.036</td>
<td>-0.461</td>
</tr>
<tr>
<td>A3</td>
<td>0.167</td>
<td>-0.340</td>
<td>0.306</td>
</tr>
<tr>
<td>A4</td>
<td>0.307</td>
<td>-0.155</td>
<td>0.171</td>
</tr>
<tr>
<td>A5</td>
<td>0.578</td>
<td>0.123</td>
<td>0.213</td>
</tr>
<tr>
<td>A6</td>
<td>0.130</td>
<td>0.518</td>
<td>0.198</td>
</tr>
<tr>
<td>A7</td>
<td>-0.457</td>
<td>-0.101</td>
<td>-0.491</td>
</tr>
<tr>
<td>A8</td>
<td>-0.229</td>
<td>-0.074</td>
<td>-0.172</td>
</tr>
<tr>
<td>A9</td>
<td>-0.102</td>
<td>-0.563</td>
<td>-0.473</td>
</tr>
<tr>
<td>A10</td>
<td>-0.102</td>
<td>-0.277</td>
<td>-0.054</td>
</tr>
<tr>
<td>A11 (GLOTTIS)</td>
<td>-0.021</td>
<td>-0.073</td>
<td>-0.004</td>
</tr>
</tbody>
</table>
Table 19. Acoustic-articulatory correlation for deaf speech.

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.118</td>
<td>-0.272</td>
<td>-0.324</td>
</tr>
<tr>
<td>A2</td>
<td>0.473</td>
<td>-0.046</td>
<td>-0.138</td>
</tr>
<tr>
<td>A3</td>
<td>0.537</td>
<td>0.231</td>
<td>0.443</td>
</tr>
<tr>
<td>A4</td>
<td>0.584</td>
<td>0.161</td>
<td>0.235</td>
</tr>
<tr>
<td>A5</td>
<td>-0.034</td>
<td>0.267</td>
<td>0.196</td>
</tr>
<tr>
<td>A6</td>
<td>-0.346</td>
<td>0.283</td>
<td>0.289</td>
</tr>
<tr>
<td>A7</td>
<td>-0.526</td>
<td>0.039</td>
<td>-0.037</td>
</tr>
<tr>
<td>A8</td>
<td>-0.661</td>
<td>-0.286</td>
<td>-0.268</td>
</tr>
<tr>
<td>A9</td>
<td>-0.421</td>
<td>-0.553</td>
<td>-0.504</td>
</tr>
<tr>
<td>A10</td>
<td>-0.089</td>
<td>-0.392</td>
<td>-0.281</td>
</tr>
<tr>
<td>A11</td>
<td>-0.012</td>
<td>-0.137</td>
<td>-0.042</td>
</tr>
</tbody>
</table>
6.1. Summary:

The goal of the present study was to investigate the possibility of using the techniques of automatic speech recognition to develop a computer system (referred to as the AIRDS system) for automatic identification and recognition of deaf speech. To achieve this goal, three objectives were identified: (1) to study the speech characteristics of deaf persons and its deviations from normal speech, (2) to identify deafness in speech, and (3) to recognize utterances produced by deaf speakers.

Identification of deafness in speech, that is, to decide whether a given speech utterance can be classified as normal or deaf speech, is treated as a speaker recognition problem in which only two speakers (or two classes of speakers) are recognized based on measurements made on the speech signal. These measurements represent four speaker dependent features, namely: intonation contours, formant transitions, timing and speaking rates, and energy contours. A distance score is computed from these measurements, and the utterance is assigned to the
class with the minimum distance. A nonlinear discriminant function rule based on the Euclidean distance is used to make this classification decision. The use of other speech features such as linear predictive coding (LPC) parameters and the vocal tract area function (VTAF) to identify deafness in speech is also investigated.

Due to the nature of deaf speech, where frequent and lengthy inter- and intra-word pauses are inserted, a two step hybrid recognizer is employed in the AIRDS system. In the first step, and before recognition, a number of candidate word boundaries are determined by marking drops in the signal energy. As a result, the input utterance is divided into a string of speech segments that could be whole words or parts of words. In the second step, each speech segment as well as all possible groupings of consecutive speech segments are matched against all reference templates. The grouping which yields the best match determines the recognition results and implicitly the word boundaries.

The AIRDS system is implemented on a PDP-11/23 minicomputer with the Interactive Laboratory System (ILS) software. A total of 468 utterances spoken by two normal hearing male speakers and two deaf male speakers are used to test the system. These utterances represent eight
phrases and 31 words which are a subset of an artificial recognition language previously designed and used for airline information and reservation tasks.

The results obtained show that measurements representing the acoustic features (intonation contours, formant transitions, timing and speaking rates, and intensity); or the linear predictor coefficients (LPC) are effective in distinguishing between normal and deaf speech. The results also show that while higher classification rates are obtained in classifying normal speech using linear predictor coefficients, similar rates are obtained in classifying both classes using the acoustic features. These rates are 87.10% and 80.65% for normal and deaf speech respectively in the first case as compared to 87.10% and 87.10% in the second case. The results also show that the classification accuracy increased successively as longer parts of the utterance were considered.

Recognition results show that higher rates were obtained in recognizing normal speech as compared to deaf speech. For isolated-word recognition, these rates are 97.85% for normal speech and 93.01% for deaf speech. For connected-word recognition, these rates are 85.61% for normal speech and 81.81% for deaf speech.
Examining the vocal tract area at the tongue and the larynx positions demonstrated that the most dramatic articulatory differences observed between the normal hearing and deaf speakers are the duration and range of movement of these articulators. These differences are most pronounced in mono-syllable words than in multi-syllable words or connected speech. Using the VTAF for speech classification, rates of 75.00% and 70.83% were obtained for normal and deaf speech, respectively.

6.2. Conclusions

While extensive research exists on developing speech training aids for deaf persons, no work accepts and uses the differences between deaf and normal speech to form templates which are typical for deaf persons with known variations. Rather, the approach to date has been in trying to make deaf speech normal before working with it.

In the present study, deaf speech is treated as having its own "norm". If this is the case, one must be able to identify and use these differences to classify speech into normal or deaf. These differences could next be used as a priori knowledge to recognize this speech.
The results of this study demonstrate that identification of deafness in speech based on measurements made on the speech signal is possible automatically. In addition, a two-step segmentation approach seems to be effective for recognition of the speech of the deaf persons. Although the classification and recognition rates for deaf speech are somewhat low when compared to normal speech, these results are encouraging for an initial system.

The results of this study also show that the vocal vocal tract area function (VTAF) could be used to estimate a set of articulatory measurements that represent the position and shape of the different articulators in the vocal tract. These measurements could be used to study the articulatory gestures of these organs, for speaker identification, or for speech recognition.

The general conclusion of this study is that the AIRDS system may be valuable in generating templates which will allow the severely handicapped deaf persons to operate and use voice-controlled machines of the future around home and work. In addition, this system could be used for developing a speech recognition-based communication aid which could be used by deaf individuals to recognize normal speech and by normal hearing
individuals to understand deaf speech. Further, this work could be extended in ways to study accent differences so that voice recognition machines will be more general and require less training by individual persons.

5.4.3. **Future Directions** :

Although the results obtained in this study are encouraging, the AIRDS system is far from complete. The possibilities for further research are many and some are suggested here.

The AIRDS system in its present form is not a real time application of speech recognition. A hardware implementation of some of the AIRDS components such as the preprocessor or the classifier will speed up the system performance. VLSI implementation of these components will be a very interesting research project.

On the other hand, further work is needed to improve the system performance. Higher rates for recognition and classification of deaf speech could be obtained by using a larger number of tokens or templates from each speaker for every word in the vocabulary, by using other distance metrics, or by using other features for speech representation.
The AIRDS system is designed with only two speakers representing each speech class. It is important, however, to test its performance on an independent data that are not used in the training. The main difficulty for this is the great variability of each utterance and the differences between speakers.

The same work could be repeated to study other speech problems such as stuttering. This will be an interesting research project and it will have a great impact in making voice recognition machines accessible by individuals with such speech problems.
REFERENCES


Davis, S.B., and Mermelstein, P. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,"


Fu, K.S. Digital Pattern Recognition, 1976.


Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. "Perception of the speech


Metz, D., Whitehead, R., and Mahshie, J. "Physiological correlates of the speech of the deaf," Deafness and Communication (D. Sims,
ed.). Baltimore, Maryland: Williams & Wilkins, 1982.


Rabiner, L.R., and Sambur, M.R. "An algorithm for determining the endpoints of isolated


Sambur, M.R., and Rabiner, L.R. "A statistical


Wakita, H. "Estimation of vocal-tract shapes from acoustical analysis of the speech wave: The state of the art," IEEE Trans., Vol. ASSP-27,


APPENDIX A

THE INTERACTIVE LABORATORY SYSTEM (ILS)

In the next few pages, an attempt is made to briefly describe the Interactive Laboratory System (ILS). For more details, the reader is invited to read the ILS documentation manuals. These include: a user guide which describes ILS commands, a training manual which assists a new user in learning how to use ILS, an installation manual which describes the installation procedure, and a programmer guide which provides information on ILS subroutines.

The Interactive Laboratory System (ILS) is a user-level application software system made up of a large number of programs for interactive signal processing. It is a modular system which teams up with a computer and graphics terminals to make up an integrated software package with the following features:

1. On-line (interactive) operation.
2. High-level language for program development.
3. Modular structure software.
4. Programs are user-oriented commands.
5. Standardized file structure.
Four types of files are defined in ILS. These are sampled data files to store integer sampled data in binary format, analysis files to store integer data from speech analysis programs, record files to store floating point data records, and label files to store information about events and their positions in a sampled data file. Each file has a header which contains information such as analysis conditions and file specifications. Files are created by the user and they could be of variable length.

Data in the analysis files are stored in vectors of length 128 locations each, with one vector for each analysis frame. Each location is reserved to store a particular data element. Some of these locations are specified as follows:

- Locations 1-30 for predictor coefficients,
- Locations 31-54 for formant frequencies, bandwidths, and amplitudes,
- Location 111 for the number of zero crossings/sec,
- Location 113 for a voicing decision,
- Location 119 for the fundamental frequency, and
- Location 123 for input signal energy.

Data in the record files are stored as feature records. Each record consists of an alphanumeric header followed by real-valued data points. These data points
are arranged as matrices with the vectors called elements and the rows called items. The header contains information describing the source of the data used to obtain that record.

The ILS commands are categorized by the type of the file processed into six functional groups. These are:

- Commands which use sampled data such as programs for elementary signal processing (ADF, CST, LSN, REC), for waveform editing and display (CUR, DSP, MDF, MVF), for spectral displays (CEP, FDI, SDI), and for digital filtering (EFI, FLT, LFI).

- Commands which use signal processing data such as programs for Fast Fourier Transform (FFT), for convolution (CNV), for correlation (COR), and for spectral density estimation (SDE).

- Commands which use speech analysis data such as programs for speech analysis (ANA, API, SIF, VTR), for speech synthesis (NSI), for speech spectral displays (FDI, CEP, SDI).

- Commands which use feature data such as programs for histograms (HIS), for scatter plots (PLR), for statistics (SME), for classification (BPA), and for template
generation (WRP).

- Commands which use label data such as programs for writing labels for specific speech events (LBF, LBA), and for listing and transferring labels (LLA, TLA).

- Commands which are general utilities such as programs for ILS initialization (ILS), for terminal selection (ASG, TRM), and for help (HHH).

The Interactive Laboratory System (ILS) supports commands to perform pattern analysis and recognition. These include commands to display data (LRE, XTR, HIS, PLR), commands to manipulate data (TRE, MRE, MDF), commands for data analysis (SME, PCO), and commands for pattern classification and recognition (BPA). Data for these commands are stored as ILS feature records in record files as described above.

REFERENCES:

ILS documentation manuals, STI, Santa Barbara, CA., 1982