


This item is held in Loughborough University's Institutional Repository (<https://dspace.lboro.ac.uk/>) and was harvested from the British Library's EThOS service (<http://www.ethos.bl.uk/>). It is made available under the following Creative Commons Licence conditions.



creative
commons


C O M M O N S D E E D


Attribution-NonCommercial-NoDerivs 2.5

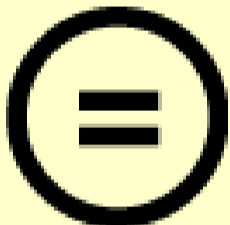
You are free:

- to copy, distribute, display, and perform the work

Under the following conditions:

 **BY:** **Attribution.** You must attribute the work in the manner specified by the author or licensor.


 **Noncommercial.** You may not use this work for commercial purposes.

 **No Derivative Works.** You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the [Legal Code \(the full license\)](#).

[Disclaimer](#) 

For the full text of this licence, please go to:
<http://creativecommons.org/licenses/by-nc-nd/2.5/>

**AN ACOUSTIC-PHONETIC APPROACH
IN AUTOMATIC ARABIC SPEECH RECOGNITION**

by

MARWAN AL-ZABIBI

A Doctoral Thesis submitted in partial fulfilment of the
requirements for the award of Doctor of Philosophy
of
Loughborough University of Technology

August 1990

Supervisor: Dr. S. Datta
Department of Electronic and Electrical Engineering

© by M AL-ZABIBI, 1990

Abstract

In a large vocabulary speech recognition system the broad phonetic classification technique is used instead of detailed phonetic analysis to overcome the variability in the acoustic realisation of utterances. The broad phonetic description of a word is used as a means of lexical access, where the lexicon is structured into sets of words sharing the same broad phonetic labelling.

This approach has been applied to a large vocabulary isolated word Arabic speech recognition system. Statistical studies have been carried out on 10,000 Arabic words (converted to phonemic form) involving different combinations of broad phonetic classes. Some particular features of the Arabic language have been exploited. The results show that vowels represent about 43% of the total number of phonemes. They also show that about 38% of the words can uniquely be represented at this level by using eight broad phonetic classes. When introducing detailed vowel identification the percentage of uniquely specified words rises to 83%. These results suggest that a fully detailed phonetic analysis of the speech signal is perhaps unnecessary.

In the adopted word recognition model, the consonants are classified into four broad phonetic classes, while the vowels are described by their phonemic form. A set of 100 words uttered by several speakers has been used to test the performance of the implemented approach.

In the implemented recognition model, three procedures have been developed, namely voiced-unvoiced-silence segmentation, vowel detection and identification, and automatic spectral transition detection between phonemes within a word. The accuracy of both the V-UV-S and vowel recognition procedures is almost perfect. A broad phonetic segmentation procedure has been implemented, which exploits information from the above mentioned three procedures. Simple phonological constraints have been used to improve the accuracy of the segmentation process. The resultant sequence of labels are used for lexical access to retrieve the word or a small set of words sharing the same broad phonetic labelling. For the case of having more than one word-candidates, a verification procedure is used to choose the most likely one.

Acknowledgements

I would like to express my sincere thanks to my supervisor Dr. S. Datta, for his advice and direction throughout this research, and also to Dr. M. Holt for his help and advice. Thanks are also due to my sponsor Scientific Studies and Research Centre (SSRC) Damascus SYRIA.

I wish to express my gratitude to Dr. M. Mrayati Director of Institute of Electronic Research at SSRC, for advice, encouragement and permission to undertake part of this research work at his laboratories. I must express my thanks to those colleagues and friends, members of the speech research group, particularly Dr. M. S. Al-Safadi, and of the computational linguistics group, particularly Mr. M. Bawwab, H. Tayyan, and Y. Meer Alam with whom I have shared many interesting and useful discussion concerning the properties of the Arabic language, and also for taking part in recording the speech database.

Finally, I would like to thank my wife Honada and my children Rim, Sa'ad, and Tariq for their patience and allowing me the time to undertake this research.

Contents

| | page |
|---|------|
| Chapter 1 Introduction | |
| 1.1 Man-Machine Communication by Speech | 1 |
| 1.2 Arabic Speech Processing | 3 |
| 1.2.1 Speech Synthesis | 3 |
| 1.2.2 Speech Recognition | 3 |
| 1.3 Database and Equipment | 4 |
| 1.3.1 Lexical Database | 5 |
| 1.3.2 Speech Database and Equipment | 5 |
| 1.4 Outline of the Thesis | 6 |
| Chapter 2 Review of Speech Recognition | |
| 2.1 Categories of Speech Recognition | 9 |
| 2.2 A Brief History of Automatic Speech Recognition | 11 |
| 2.3 Isolated Word Speech Recognition | 14 |
| 2.3.1 Feature Measurements | 15 |
| 2.3.2 Pattern Matching Model | 21 |
| 2.3.3 Recognition Without Time Alignment | 31 |
| 2.3.4 Hidden Markov Modelling | 32 |
| 2.3.5 Neural Network | 36 |
| 2.3.6 Acoustic Phonetic Approach | 37 |
| 2.3.7 Syntactic Pattern Recognition | 41 |
| 2.4 Connected Word Recognition | 42 |
| 2.5 Continuous Speech Recognition | 45 |
| 2.5.1 Speech Understanding Concept | 45 |
| 2.5.2 Sources of Knowledge | 46 |
| 2.5.3 Speech Understanding Models | 49 |
| 2.5.4 Remarks on Continuous Speech Recognition | 51 |
| 2.6 Word Spotting | 52 |
| 2.7 Summary | 53 |

Chapter 3 Arabic Phonetic System, Phonology and Morphology

| | | |
|-------|--|----|
| 3.1 | Introduction | 54 |
| 3.2 | The Human Vocal Mechanism | 54 |
| 3.3 | Articulatory Phonetics | 56 |
| 3.4 | Vowels | 59 |
| 3.5 | Consonants | 61 |
| 3.5.1 | Consonant Classes | 63 |
| 3.5.2 | Vowels in Pharyngealised Context | 64 |
| 3.6 | Distribution of Vowels and Consonant Clusters | 68 |
| 3.6.1 | Distribution of Vowels | 68 |
| 3.6.2 | Distribution of Consonant Classes | 69 |
| 3.7 | Syllabic Types and Structure (Phonology) | 71 |
| 3.7.1 | Syllabic Types | 71 |
| 3.7.2 | Phonological Constraints | 72 |
| 3.7.3 | Gemination | 73 |
| 3.7.4 | Distribution of Syllabic Patterns | 75 |
| 3.8 | A Brief Description of the Arabic Morphological System | 77 |
| 3.9 | Prosodic Features | 80 |
| 3.9.1 | Stress | 80 |
| 3.9.2 | Duration | 81 |
| 3.9.3 | Intonation | 82 |
| 3.10 | Summary | 82 |

Chapter 4 A Model of Lexical Access for a Large Vocabulary Recognition System

| | | |
|-------|---|----|
| 4.1 | Introduction | 83 |
| 4.2 | Broad Phonetic Classification and Lexical Access | 84 |
| 4.3 | Discrimination of Words | 86 |
| 4.3.1 | Classification Schemes | 88 |
| 4.3.2 | Statistical Results | 89 |
| 4.3.3 | Effect of Syllabic Structure on the Classification Results... | 91 |
| 4.3.4 | Discussion | 94 |
| 4.4 | Structure of the Lexicon | 95 |

| | | |
|-----|---|----|
| 4.5 | An Automatic Speech Recognition Model | 96 |
| 4.6 | Speech Recognition Experiment | 97 |
| 4.7 | Summary..... | 99 |

Chapter 5 Preliminary Segmentation and Vowel Recognition

| | | |
|-------|--|-----|
| 5.1 | Introduction | 100 |
| 5.2 | Voiced-Unvoiced-Silence Segmentation | 101 |
| 5.2.1 | Initial Silence Detection | 103 |
| 5.2.2 | Pitch Detection | 103 |
| 5.2.3 | Initial Voiced-Unvoiced Classification | 104 |
| 5.2.4 | Weak Fricative Detection | 108 |
| 5.2.5 | Editing V-UV-S Contour | 108 |
| 5.3 | Vowel Detection | 113 |
| 5.3.1 | Energy Peak Detection | 113 |
| 5.3.2 | Loudness Peak Detection | 124 |
| 5.3.3 | Results of Vowel Detection | 125 |
| 5.4 | Vowel Identification | 128 |
| 5.4.1 | Vector Quantisation Method | 128 |
| 5.4.2 | Formant Method | 133 |
| 5.4.3 | Vowel in Pharyngealised Context | 140 |
| 5.5 | Summary | 142 |

Chapter 6 Spectral Variation Contour and its Application to Speech Segmentation

| | | |
|-------|---|-----|
| 6.1 | Introduction | 145 |
| 6.2 | Spectral Variation Contour | 146 |
| 6.3 | Parametric Representation | 147 |
| 6.3.1 | Cepstral Parameters | 147 |
| 6.3.2 | Mel-Frequency Cepstral Parameters | 149 |
| 6.3.3 | MFCCs Computation | 152 |
| 6.4 | Transition Detection and Segmentation | 157 |
| 6.5 | Summary | 169 |

Chapter 7 Segmentation and Error Correction

| | | |
|-----|--|-----|
| 7.1 | Introduction | 171 |
| 7.2 | Plosive Detection | 172 |
| 7.3 | Segmentation and Labelling | 174 |
| | 7.3.1 Endpoint Adjustment | 175 |
| | 7.3.2 Labelling | 178 |
| 7.4 | Error Correction Procedure | 195 |
| | 7.4.1 Detection of Geminated Consonant | 196 |
| | 7.4.2 Plosive and Glottal Correction | 200 |
| | 7.4.3 Vowel Correction | 203 |
| | 7.4.4 Syllabic Pattern Correction | 210 |
| 7.5 | Results and Discussion | 221 |
| 7.6 | Summary | 224 |

**Chapter 8 Conclusions and
Suggestions for Further Work**

| | | |
|-----|--|-----|
| 8.1 | Introduction..... | 225 |
| 8.2 | Phonetic Classification Schemes | 226 |
| 8.3 | The Word Recognition Model | 227 |
| 8.4 | Further Consonant Classification | 229 |
| 8.5 | Closing Remarks | 231 |

| | |
|-------------------------|-----|
| References | 232 |
|-------------------------|-----|

Appendices

| | | |
|----|--|-----|
| A. | A List of the words in the speech database | 247 |
| B. | Spectrograms of the Consonant-Vowel Pairs | 253 |
| C. | Mel Scale and Critical Bands | 261 |

Chapter 1

Introduction

1.1 Man-Machine communication by Speech

Speech communication between man and machine introduces a new range of communication services which extend man's capabilities, serve his social needs and increase his productivity.

Communication between people and machine includes automatic speech recognition and automatic speech synthesis. In speech synthesis, the machine or computer takes the speaker's role in generating speech from a pre-defined textual message, while in speech recognition, the machine takes the listener's role in decoding speech waves into either the underlying textual message or a hypothesis concerning the speaker's identity.

The current advances in the microelectronic devices (semi-conductor technology) with the advent of digital signal processors (DSP), have facilitated the availability of complex commercial speech synthesis and recognition systems.

Speech synthesis is a reasonably well established field if measured in terms of the range of products currently in the market. Speech recognition is not so well developed as it is inherently a much more difficult problem because of asymmetries in producing and interpreting speech.

The design and implementation of voice interaction between man and machine requires the involvement of a wide spectrum of disciplines, namely linguistics (i.e., phonetics, phonology, prosody, and computational linguistics), computer science, ergonomics, and speech processing. Therefore unrestricted speech synthesis or recognition is considered as a language-dependent problem.

The likely areas of application for man-machine speech communication are growing rapidly. Such applications are listed below, where some of them are already commercially available, especially the speech synthesis products.

a. Automatic information services

These are handled by making a speech link with a computer (e.g., through a telephone line) and retrieving the requested information such as enquiries, reservations, bank services, and computerised data banks.

b. Consumer products such as:

- Talking clocks, calculators, toys, warning systems and voice controlled devices.
- Office automation systems such as automatic typewriter.
- In quality control tasks, automatic material handling and stock control, where the operator's hands and eyes are fully occupied.

c. Services for the handicapped such as:

- Blind students can have access to computer assisted learning methods through speech synthesis.
- Physically disabled persons who can not manipulate the buttons of a computer's keyboard or any other devices, can use the computer through speech recognition.

d. Security applications such as:

- Speaker verification based on speech can be used along with magnetic card or badge reader to control entry to restricted areas such as classified record storage, classified research laboratory, etc.
- Surveillance of communication channels, where listening to radio broadcasts or any other narrow-bandwidth communication media is a time-consuming, manpower-intensive, tedious task for operators. A potential solution to this problem is the use of automatic speech recognition technology to automate part of the listening process. This process can include:
 - Message sorting through speaker identification.
 - Word spotting, to recognise keyword or a set of keywords embedded in the conversational speech.
 - Language identification.

1.2 Arabic Speech Processing

Speech processing for English, French, and many other languages has been the object of much research in the last thirty years. Some of its problems have been successfully solved, while others are still under research.

Text and speech processing for the Arabic language has been the object of much research in the past ten years throughout the Arab world [1]. Research on Arabic speech processing is growing slowly [2,3], while research on Arabic text processing (e.g., text compression, computer aided translation, and natural language understanding) is growing rapidly. Research on Arabic speech synthesis is advancing faster than that on speech recognition, perhaps due to the inherent difficulties of the speech recognition task, and the lack of modern studies on acoustic-phonetic, phonemics, and prosodic of Arabic.

1.2.1 Speech Synthesis

Research on Arabic speech synthesis has been carried out in many Arab countries, especially at the Faculty of Sciences of Rabat University, MOROCCO, at the Kuwait Institute for Scientific Research (KISR), KUWAIT, and at the Scientific Studies and Research Centre (SSRC), Damascus, SYRIA. Apart from vocoding techniques, various synthesis techniques have been employed such as:

- Synthesis by pre-analysed words [2].
- Synthesis by diphones [2,4].
- Synthesis by sub-syllabic sound units [5].

Recently research work is going on at SSRC to achieve text to speech synthesis using articulatory parameters and rules, by exploiting a newly developed speech production theory [6].

1.2.2 Speech Recognition

A few research works have been done on automatic Arabic speech recognition. Pattern matching and hidden Markov modelling have been used to recognise a small set of isolated words (e.g., the 10 digits, 8 words to control a wheelchair, and vowel identification), [7,8,9]. These methods are actually language-independent methods.

Therefore the lack of any particular research relating to Arabic speech recognition which exploits features of the Arabic language, is actually the main motivation of this research.

Small vocabulary recognition systems in use today are limited in usefulness by the restrictions they impose on the human user. Large vocabulary systems capable of recognising thousands of words could meet many of the needs specified by the above mentioned applications. In addition to the capability of handling a large vocabulary, the speech recognition system must be capable of easily learning new words and updating its internal database. It is also preferable that the system is capable of recognising words spoken in continuous sentences, and is speaker independent.

This research is mainly concerned with a large vocabulary isolated word Arabic speech recognition, to fulfil much of the above requirements. Research on isolated words was chosen to simplify the recognition process, where it could be considered as a first step towards further research on continuous speech recognition. Most of the work in this thesis, such as the phonetic studies, and the developed recognition techniques, can be carried over into continuous speech recognition.

An acoustic-phonetic approach is employed in this research work. In this approach, detailed vowel recognition is performed, while consonant recognition is achieved according to broad phonetic classes. The broad phonetic analysis of consonants is chosen to overcome the variabilities in the acoustic realisation of utterances. Statistical studies are carried out on a large vocabulary to investigate the effectiveness of this approach.

In order to tackle a large vocabulary, a knowledge of the Arabic phonetic system is essential. Also, phonological and morphological knowledge are required in addition to the speech processing techniques to facilitate and enforce the recognition process. Therefore the Arabic phonetic system, phonology, and morphology of the Arabic language, are presented in this thesis.

1.3 Database and Equipment

Two databases have been used in this research work, which are a lexical database and a speech database.

1.3.1 Lexical Database

In order to justify the use of detailed vowel recognition and broad phonetic consonant classification to discriminate between words in a large vocabulary, statistical studies have been carried out on a large vocabulary database. These studies also demonstrate the distribution of vowels, different consonant classes, and different syllabic patterns.

The database has to satisfy two conditions. Firstly, it has to comprise a selection of words that are commonly used in natural language (speech). Secondly, it has to include various kinds of words and their derivatives (which are useful in the actual recognition process to cover all possible patterns).

To comply with the above conditions, two lexicons have been chosen and included in this database.

The first lexicon comprises the most common (frequently used) 3,000 Arabic words reported in the literature [10]. These words were extracted from about one million words, and represent about 84% of the total words. The size of the words in this lexicon varies from 1 to 4 syllables, where the most frequent words are actually the shorter they are (in terms of number of syllables).

The second lexicon comprises 10,000 randomly chosen words. However, it includes almost all the words of the first lexicon besides other polysyllabic words (up to 7 syllables). Acoustically similar words and several derivatives of several words (according to different morphological patterns) are also included in this lexicon.

The two lexicons are stored into data files. An orthographic to phonemic procedure (programme) has been implemented to transfer the words in the two lexicons into phonemic form according to the standard Arabic pronunciation (which is used in official discourse, teaching, and literature throughout the Arab world).

1.3.2 Speech Database and Equipment

A speech database is used for the development and testing of all the recognition algorithms presented in this thesis. The aim of the speech recognition experiment, which uses this database, is to demonstrate the effectiveness of the proposed classification

schemes, where a novel segmentation procedure has been developed during the research work.

The speech database consists of a set of 100 words uttered mainly by three cooperative male speakers. In addition to that, some words (up to 50 words of this set) uttered by various speakers (male, and female) are also tested.

The speech data were recorded in different environments (i.e., in a laboratory room, in an office, and in a house), at different time intervals, using a commercial cassette tape recorder, and a head mounted noise-cancelling omnidirectional dynamic microphone (Shure SM10A). The speech signal was low-pass filtered to 10 KHz, sampled at 20 KHz, digitised through a 12 bit A/D converter and stored on disk. All speech data were also digitally filtered to 4.8 KHz, down-sampled to 10 KHz, and then stored into data files for later processing.

A Digital Sona-Graph has been used to make spectrograms for all the words in the speech database, and for all possible combinations of consonant-vowel pairs. The spectrograms were made with a broad-band analysis filter (300 Hz), which provides accurate timing resolution. The spectrograms were made by using the Kay Sona-Graph model 7800.

The results presented in this thesis were obtained through simulation on a VAX-11 780 under the VMS 3.4 operating system.

1.4 Outline of the Thesis

This thesis presents a research work on a large vocabulary isolated Arabic words speech recognition using an acoustic-phonetic approach. The thesis is organised as follows:

Chapter 2 presents a review of speech recognition categories and techniques. It first introduces the various speech recognition categories. Then it demonstrates most of the methods used for isolated word recognition, such as pattern matching using dynamic time warping, vector quantisation, Markov modelling, and acoustic phonetic approaches. In this context, various acoustic parameters employed in these methods are presented. It also demonstrates some approaches for connected speech recognition, continuous speech recognition, and speech understanding. The chapter ends with a brief description of some techniques used in word spotting.

Chapter 3 describes some of the linguistic aspects of the Arabic language which are necessary for this research. The elements of the phonetic system (vowels and consonants) are presented, and the pharyngealisation phenomenon related to some consonants is demonstrated. Distributions of vowels and consonant clusters using the lexical database are reported. This chapter also discusses the syllabic types and patterns of the Arabic language and reveals some phonological constraints. It ends with a brief description of the Arabic morphological system.

A model of lexical access for a large vocabulary recognition system is introduced in Chapter 4. Word discrimination in a large vocabulary using phonetic description is investigated, where different phonetic classification schemes are used. The results of the classification schemes using two lexicons are reported. This chapter demonstrates the effect of detailed vowel recognition on the results of broad phonetic classifications. Next, a proposal on how words can be structured in the lexicon of a recognition system is introduced. The chapter ends with an outline of the proposed speech recognition model which has been used in the recognition experiments throughout this work.

Two main procedures are introduced in Chapter 5. These are voice-unvoiced-silence segmentation and vowel recognition. In the former procedure, the speech signal is segmented reliably into voiced speech, unvoiced speech, and silence (no speech). The vowel recognition procedure consists of two main phases, i.e., vowel detection and vowel identification. The vowel detection phase describes how vowels are located, while the identification phase determines vowel identities. Two different methods for vowel identification are demonstrated in this chapter. These methods are vector quantisation, and the formant method. The chapter ends by demonstrating some cues related to vowels in pharyngealized consonantal context.

Chapter 6 describes the development of an automatic procedure for detecting transition between adjacent phonemes. It starts with modelling the speech signal in terms of mel frequency cepstral parameters. Then, it describes the computation of the spectral variation contour along a word. This contour determines the transitional as well as the steady-state regions along the spectrum of a certain word. This chapter discusses the usefulness of the spectral variation contour for the segmentation process which is carried out in the following chapter.

In Chapter 7, the results of the voiced-unvoiced-silence segmentation, the vowel recognition, and the transition detection, are employed to perform the broad phonetic segmentation and labelling process. The resultant string of labels which describes a certain input word is passed through an error correction procedure. This procedure tackles all sorts of expected errors concerning vowels, consonants, and the syllabic pattern as a whole for the input word, by utilising durational information and phonological constraints.

The final chapter provides a recapitulation of both the novel recognition scheme proposed in this thesis and the main results obtained experimentally by computer simulation. Suggestions for further research works are also included in this chapter.

Chapter 2

Review of Speech Recognition

2.1 Categories of Speech Recognition

Automatic speech Recognition tasks can be classified according to the following categories:

- Isolated word recognition
- Connected word recognition
- Continuous speech recognition
- Speech understanding
- Word spotting
- Speaker identification and verification
- Language identification

In isolated word recognition, the words are spoken in isolation. Pauses between words simplify recognition because they make it relatively easy to identify endpoints (i.e., the start and end of each word), and they minimise coarticulation effects between words. In addition, isolated words tend to be pronounced somewhat more carefully, since the need to pause between words impedes fluency, which would otherwise tend to encourage a more natural and hence more careless pronunciation. Isolated words are adequate for many applications but are far from being a natural way of communication.

In connected word recognition, the spoken input is a sequence of isolated words from a specified vocabulary and the recognition is based on recognising isolated words.

The recognition of continuous speech is an attempt to transcribe naturally spoken utterances (i.e., without artificial pauses between phonemes, syllables, words, or sentences) in accordance with the rules of language orthography. This implies the need for some form of segmentation of the speech into linguistic units. The fluency of speech in natural speech imposes co-articulation between adjacent phonemes and words in a

phrase. This leads to neglecting some phonemes in a phrase, especially between words, which makes the recognition process very difficult to achieve.

The goal of a speech understanding system is to identify the meaning of the speech without constraining the speaker's sentence structure. In such a system, traditional speech recognition techniques are integrated with artificial intelligence techniques to give the extra power needed to deal with natural continuous speech. High-level knowledge sources (i.e., morphological, syntactic, semantic, and pragmatic) are incorporated in this system.

In word spotting, the speech recognition deals with detecting the occurrence of a given word in continuous speech. In this case, all the speech is ignored until a keyword is spoken. Therefore the system is tuned to recognise words which have high correlation to one of the pre-specified keywords.

In speaker identification and verification, the aim of the speech recognition here is not to recognise what has been said but actually to highlight differences between speakers. In speaker identification, an unknown speaker is to be recognised from a previously specified group of speakers, while in speaker verification, the speech recognition technique is used in addition to other identification systems (such as a magnetic card reader) to verify the identity of the speaker.

In language identification, the speech understanding techniques are used to form some sort of linguistic chains from the phonetic transcription of speech, and these are used as a means of discrimination between different languages.

Two terms which are frequently used to describe a speech recognition system are speaker-dependent and speaker-independent. In a speaker-dependent system, the system is to be trained to the speech of each new speaker for the entire vocabulary. In a speaker-independent or multi-speaker system, no training is required for the new speaker. Actually, for a large vocabulary system and for continuous speech recognition, instead of full training the system can adapt to a new speaker by some relatively simple restricted procedures using a few words or sentences. The latter case is often called speaker adaptation.

In the following sections, a brief history of automatic speech recognition and some of the

techniques used in isolated word recognition are presented. Then a brief description of connected speech, continuous speech and speech understanding systems are given. This chapter ends with a brief introduction to word spotting. Speaker identification and verification will not be discussed in this thesis.

2.2 A Brief History of Automatic Speech Recognition

Man has always been fascinated by his ability to speak. Without this advanced way of communication the establishment of society as we know it would have been impossible. Ideas would have not been readily communicated and man's superiority over other animals would have been diminished. Attempts were made to model the human speech production model two centuries ago [11].

a. The Early Work of the Pre-Sixties:

The introduction of the vocoder by Dudley in 1939 [12], and of the sound spectrograph in 1947 [13], gave a better understanding of the information-bearing elements in a speech signal. The spectrograph had shown that different spoken words gave rise to different acoustic patterns. It was therefore believed that all the information required for recognising speech resided in the acoustic signal.

The first attempt to build a recogniser based on acoustic patterns was by Davis *et al* at Bell laboratories [14]. They devised an apparatus for recognising digits spoken in isolation. Their method of analysis was based on dividing the frequency spectrum of the speech signal into two bands, one above and one below 1000 Hz. The number of zero-crossings in each band was then counted, giving an approximate measure of the first and second formants frequencies. A matrix with 30 elements representing the F1-F2 plane was thus established. A reference pattern was formed for each digit using the F1-F2 trajectory in this matrix. When a new digit was spoken the pattern produced was cross-correlated with each of the stored reference patterns. This gave an approximate measure of the probability that a particular digit had been spoken, and so enabled the most likely digit to be chosen. Provided that the reference patterns were adjusted for a particular speaker, it was reported that the digit spoken was correctly recognised in about 98% of cases. With a new speaker, however, with no adjustments, the recognition score was often as low as 50%.

Apart from its historical significance, the Davis recogniser introduced the technique of reducing the input speech signal into a pattern and then comparing it with pre-stored reference patterns, a method which is in force today.

Dudley and Balashek [15] developed a machine which performed a spectral analysis of the speech signal with a bank of ten band-pass filters each 300 Hz wide. The output of the filter bank was cross-correlated with stored patterns of spoken digits, and the best match was selected. This system also produced good results with the speaker who generated the patterns, but was less successful with other speakers.

Another early attempt at automatic speech recognition was the so-called 'phonetic typewriter' of Olson and Belar [16]. This system also used a bank of 8 filters, but also employed a compressor which attempted to adjust the mean level of the signal to about the same intensity for both quiet and loud speakers to reduce some of the variability. The outputs from the filter bank set relays every 40 msec if a threshold current had been exceeded. The output was thus a crude spectrogram, which was decoded as one of the ten syllables (used in the system), and hence into individual letters which actuated typewriter keys. The machine was tested with sentences consisting of the pre-defined set of 10 syllables in various permutations. With careful pronunciation it was claimed that an accuracy of 99% was obtained.

A system based on distinctive feature theory was developed by Wiren and Stubbs [17]. In this recogniser, a binary classification was used. The voiced sound was separated from voiceless. The voiceless sounds were then divided into fricatives and plosives. The binary classification was repeated until a single phoneme was isolated. The decisions were based on the acoustic features presented in the signal. Fairly good results were achieved by this system.

The overall performance of these early recognisers, especially in a speaker-independent mode, was not impressive. Nevertheless, These early attempts at recognition did demonstrate the value of using the spectrograph as a useful tool in speech recognition.

b. The Work in the Sixties:

The use of the digital computer in speech recognition was first employed in the early 1960's. One of the earliest speech recognition systems to deal successfully with a number

of different speakers was that of Forgie and Forgie [18]. They noted that there is an inverse relationship between the pitch of the voice and the size of the vocal tract, so the fundamental frequency can be used to normalise formant frequencies. Their speech recogniser consisted of a 35-channel filter bank connected to a computer. The formant frequencies of an incoming speech signal were determined from the spectrum, and then normalised. From these measurements the vowels in the context /b/-vowel-/t/ words were recognised. They reported an accuracy of 93% for ten vowels spoken by each of 21 male and female speakers with no adjustment for the speakers.

A similar speech recogniser was built by Denes and Mathews [19], where a 17-channel spectrum analyser formed the input to a computer. Spectrum patterns were formed from the spoken digits, and a number of utterances of each word were averaged and stored as reference patterns. Unknown utterances were recognised by comparing them with the stored patterns by a cross-correlation process. The novel feature of this system was that a provision was made for the duration of the patterns to be normalised before classification. The system was tested with one female and six male speakers. An error rate of 6% was obtained with normalisation, and 12% without it.

Sakai and Doshita reported a more comprehensive recogniser [20]. They used separate circuits for segmenting the speech into vowels and consonants and for classifying the segmented phonemes. Zero-crossing analysis was combined with measurements of the variation of energy in various frequency regions. They claimed 70% correct recognition on consonants and 90% on vowels, although it was pointed out that some phonemes were not allowed as input.

The introduction of the Fast Fourier Transformation (FFT) in the mid sixties by Cooley and Tukey [21], made it possible to achieve complex mathematical analysis of speech waveforms with reasonable computational effort and also paved the way for fully digital speech recognition systems. This, along with the desire to market small scale recognition products, led to the development of special purpose hardware.

c. The Work in the Early Seventies:

At the end of the sixties and beginning of the seventies, speech scientists had begun to expand their domain to the recognition of continuous speech.

There are many sources of linguistic knowledge which may be used in a speech recogniser in order to improve its performance. In the early seventies it was felt that the use of such sources of knowledge could be exploited to solve the problem of recognising spoken sentences and longer periods of speech. In the USA a major effort, funded by the Advanced Research Projects Agency (ARPA), was commenced, to tackle the problems encountered in continuous speech, as opposed to isolated word recognition. A five-year research programme was initiated [22], where different sources of knowledge were integrated into a network. The ARPA project called for a system that would accept continuous speech from any cooperative speaker. The language was limited to a vocabulary of 1000 words and was allowed to have an artificial syntax appropriate to a limited task situation, e.g., data management, chess playing, etc.

When the ARPA project ended in 1976, a number of task-dependent systems: HARPY [23], HEARSAY [24], HWIM [25], which could understand spoken utterances within a given context, had been developed. Many of the present day continuous speech recognition systems still employ the techniques investigated during the ARPA project.

Some speech recognition systems had used a probabilistic function of a Markov process to model the speech signal [26]. Other algorithms based on stochastic modelling, to model the linguistic knowledge sources necessary for continuous speech recognition, gave encouraging results [27].

In conjunction with the above mentioned systems, two other major developments in the early seventies had helped to accelerate the recognition research, especially for isolated words. These were the introduction of linear prediction coding (LPC) and dynamic time warping (DTW) techniques, as we will see later on in this chapter.

Further developments in the speech recognition in the mid seventies and during the eighties will be reported in the following sections, where various techniques used in the recognition of isolated and continuous speech recognition will be presented.

2.3 Isolated Word Speech Recognition

Isolated word speech recognition can be dealt with using two main approaches: the mathematical approach and the acoustic-phonetic approach. In the former approach, pattern matching methods, stochastic modelling using hidden Markov models, and more

recently neural networks, have been employed. These methods utilise little or no speech specific knowledge. The acoustic-phonetic approach utilises linguistic knowledge such as: phonetic, phonological, and morphological knowledge. Descriptions of these different techniques are introduced in the following sections, but before going into that, the acoustic parameters used in most of the speech recognition methods are described in the next section.

2.3.1 Feature Measurement

A speech signal is a highly redundant signal. It carries linguistic messages as well as other information about speakers, regarding their physiology, psychology, etc. Feature measurement, some times called feature extraction, is basically a data reduction technique. The digitised speech signal is transformed into a smaller set of features which faithfully describe the salient properties of the acoustic waveform. Data reduction rates (or compression ratios) of 10 to 100 are generally practical.

A number of different feature sets have been proposed ranging from simple sets such as energy and zero-crossing rates to complex representation such as:

- Short-time spectrum (DFT or filter bank)
- Linear predictive coding
- Cepstral parameters (homomorphic model)
- Articulatory parameters
- Auditory model

The motivation for choosing one feature set over another is often dependent on the constraints imposed on the system in terms of cost, speed, and recognition accuracy.

Before we discuss some of the feature sets used in the speech recognition systems, a brief description of the speech production model is introduced.

a. Speech Production Model

In the speech production model, the speech signal is modelled as the output of a linear time-varying system excited by either quasi-periodic pulses (for voiced sounds), or a random noise signal (for voiceless or unvoiced sounds) as illustrated in Figure 2.1. The

linear time-varying system represents the vocal tract, while the periodic pulses represent the vocal cord vibration, and the random noise represents the air turbulence during uttering voiceless sound [28]. The vocal tract could be considered as a linear system for a short time, during which the speech signal is considered as stationary, while it is non-stationary over a long time duration. From Figure 2.1 we can write

$$s(t) = g(t) * h(t) \quad (2.1)$$

where $s(t)$ is the speech signal, $g(t)$ is the source signal, and $h(t)$ is the impulse response of the vocal tract. In the Frequency domain, Eq. (2.1) becomes:

$$S(F) = G(f) \cdot H(f) \quad (2.2)$$

where $H(f)$ represents the transfer function of the vocal tract filter or the spectral envelope of the speech signal. The resonances arising in this envelope are referred to as 'Formant frequencies' or simply 'Formants'. Figure 2.1 shows the signals of the speech production model in the time and frequency domains for two cases; voiced sound and voiceless sound. The vocal tract takes different shapes when pronouncing different phonemes, and this actually leads to different transfer functions (spectral envelopes), and hence to different formant frequencies. It is usual to find at least three formants below 4 KHz.

The source or excitation is typically represented in terms of the voicing decision, the overall amplitude, and the fundamental frequency estimate (F_0). Spectral information is weighted much more heavily than the excitation data in speech recognition because amplitude and F_0 are more influenced by higher-level linguistic phenomena rather than by phonemics. Most recognisers use a set of features which model the spectral envelope as we will see later on.

To provide an efficient feature representation of speech, three or four formants are considered sufficient to model the spectral behaviour of the short time spectral envelope. Usually the four formants are found below 4 KHz. Spectral details at frequencies above F_4 contain phonemic information (e.g., equal-spaced harmonics indicate voiced speech, high energy there suggests fricative sound). A wider bandwidth of up to 6.4 KHz can be used to improve the recognition of some fricative consonants.

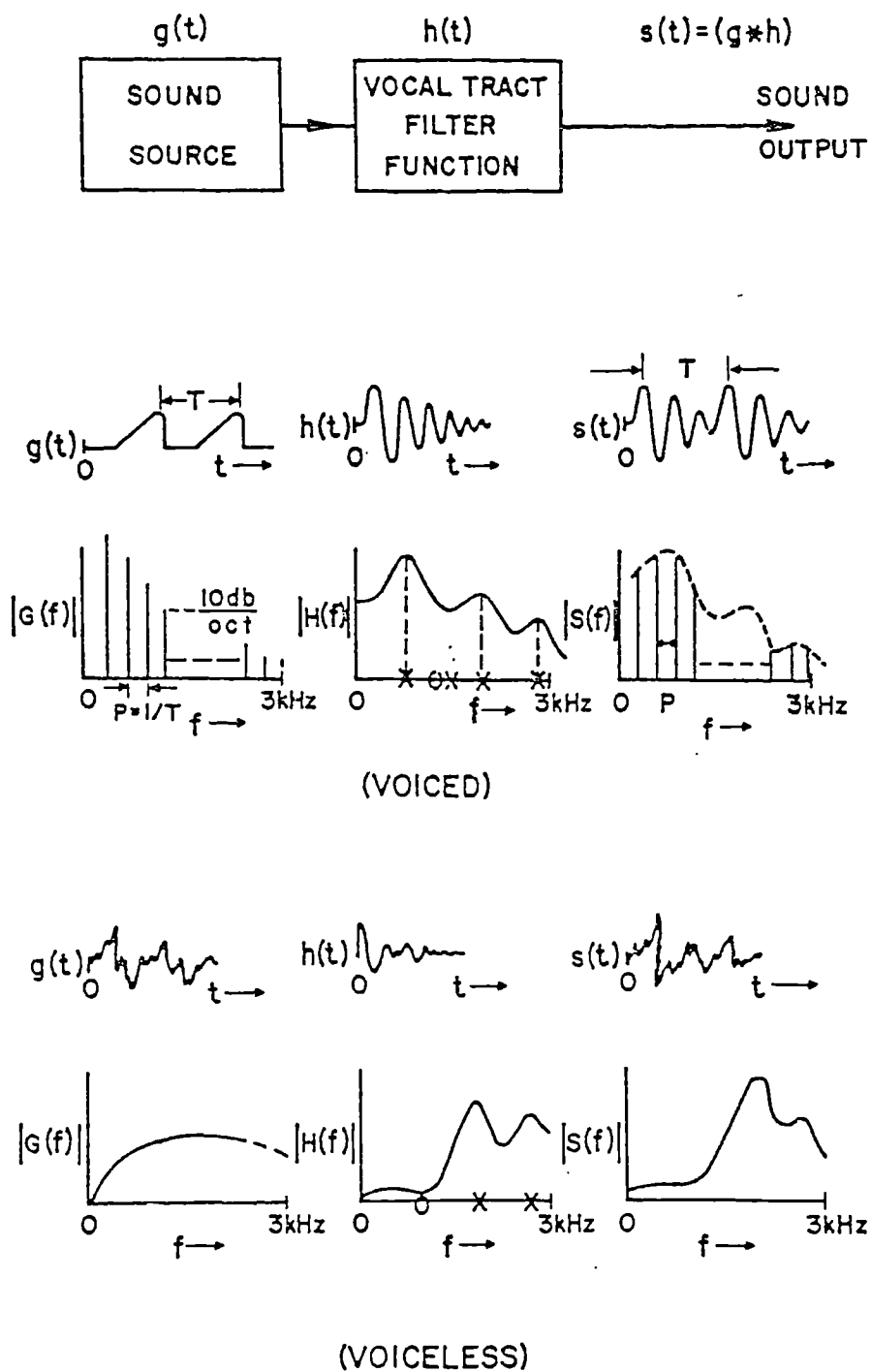


Fig. 2.1 Speech production model (source-vocal tract)

b. Linear Prediction Parameters

Linear prediction coding coefficients can model the spectral envelope well, and are widely used. The basic idea behind LPC is that a given speech sample can be approximated as a linear combination of past speech samples [29]. For each sample a prediction error $e(n)$, is defined as follows:

$$e(n) = s(n) - \bar{s}(n) \quad (2.3)$$

$$\bar{s}(n) = \sum_{i=1}^P a(i) s(n-i) \quad (2.4)$$

$$H(z) = \frac{1}{1 - \sum_{i=1}^P a(i) z^{-i}} \quad (2.5)$$

where $\bar{s}(n)$ is the linearly predicted sample, $s(n)$ is the actual sample, P is the degree of the LPC model (filter), and $a(i)$ where $i=1,2,\dots,P$ are the filter predictor coefficients. By minimising the mean-square prediction error $e(n)$, over a finite interval, a unique set of predictor coefficients can be determined. The LPC coefficients give good short-time spectral estimation of the linear time varying system. $H(z)$ in Eq. (2.5) represents the z -transform of the transfer function of the vocal tract (all pole model).

For a short interval (M samples of speech), the LPC coefficients are computed to yield an N -dimensional feature vector, where N equals P (the model's degree) which is usually taken between 8 to 14 [30]. The time variation of these feature vectors defines a pattern for the speech utterance.

Formant frequencies and their bandwidths can be extracted from the transfer function of the vocal tract by a peak picking procedure. Computing the FFT over the set of LPC parameters and taking the inverse of the result, yields the transfer function of the vocal tract (Eq. (2.5)). Another way to find the formant frequencies and their bandwidths is to solve the inverse of Eq. (2.5) and find its roots (complex pole-pairs) [30].

c. Filter Bank Parameters

A popular set of features used in many speech recognition systems is the output of a bank of filters. The speech signal is passed through a bank of bandpass filters covering the speech bandwidth. The energy at the output of each channel is estimated from the output of each particular filter [32]. The set of energy values at each interval of time (frame) constitutes an N-dimensional feature vector. The time variation of these feature vectors defines a pattern for the speech utterance. In general the bandpass filters are linearly spaced at low frequencies (below 1000 Hz) and logarithmically spaced at high frequencies. It was found [33], that 13 filters spaced along a critical-band frequency scale (or bark scale), were enough for high recognition accuracy, and using 15 filters spaced uniformly in frequency gave the same result as critical-band filters in a template matching approach.

d. Cepstral Parameters

Three types of cepstral parameters have been used in speech recognition systems (homomorphic model) [34], namely the linear frequency cepstral coefficient (LFCC) [35], the mel-frequency cepstral coefficients (MFCC) [36], and the LPC-derived cepstral coefficients (LPCC) [31].

The LFCCs are computed from the log-magnitude discrete Fourier transform (DFT) directly as follows:

$$\text{LFCC}_i = \sum_{k=0}^{K-1} Y_k \cos\left(\frac{\pi i k}{K}\right) \quad (2.6)$$

where $i = 1, 2, \dots, N$. K is the number of DFT log-magnitude coefficients Y_k , and N is the number of employed cepstral coefficients.

In mel-frequency scale, the DFT magnitude spectrum is frequency-warped to follow a critical band scale (mel-scale) [36, 37] and amplitude-warped (logarithmic scale), before computing the inverse DFT parameters. Therefore Q bandpass filters are used to cover the required frequency range, and the MFCCs are computed as follows:

$$\text{MFCC}_i = \sum_{k=1}^Q X_k \cos \left[i \left(k - \frac{1}{2} \right) \frac{\pi}{Q} \right] \quad (2.7)$$

where $i = 1, 2, \dots, N$. N is number of cepstral coefficients used, and X_k represents the log-energy output of the k th filter.

The LPCCs are obtained from the LPC parameters directly as follows:

$$\text{LPCC}_i = \text{LPC}_i + \sum_{k=1}^{i-1} \frac{k-i}{i} \text{LPCC}_{i-1} \quad \text{LPC}_k \quad (2.8)$$

where $i = 1, 2, \dots, N$. N is number of employed coefficients. For i greater than the order of the LPC model, LPC_i is taken equal to zero.

The set of N parameters (LFCCs, MFCCs, or LPCCs) constitutes an N -dimensional feature vector. The time variation of these feature vectors defines a pattern for the speech utterance. It was found that 6 MFCCs gave better accuracy than any other 10 (or more) cepstral coefficients [35].

e. Articulatory Parameters

Another set of features for describing speech sounds would be the parameters giving the position of the tongue, lips, jaws and the velum as functions of time. These parameters can be estimated from the speech signal [38]. A new speech production theory based on distinctive regions along the vocal tract has been introduced [6, 39], which provides a new concept in the acoustic-articulatory-phonetic relation. By performing acoustic-articulatory inversion, the area function can be used as an articulatory parameter for speech recognition.

f. Auditory Model Parameters

Another approach for feature measurements is the use of the auditory model [40]. The psychophysical aspects of critical bandwidth, loudness, timbre, and subjective duration have been used as feature measures [41]. Another design which tries to capture the

time-varying nature of the auditory model by combining the psychophysical critical-band, and loudness estimation with a firing-rate model, has improved the accuracy of the speech recognition compared to previous filter-bank feature measures [42].

2.3.2 Pattern Matching Model

The classical way of solving isolated word recognition problem is to treat it as a pattern recognition problem, where digital signal processing techniques can be applied to obtain a pattern for each word.

Figure 2.2 displays the typical pattern matching model employed in the majority of isolated word recognition systems. This model consists of three stages:

- Feature measurement
- Pattern comparison
- Decision rule

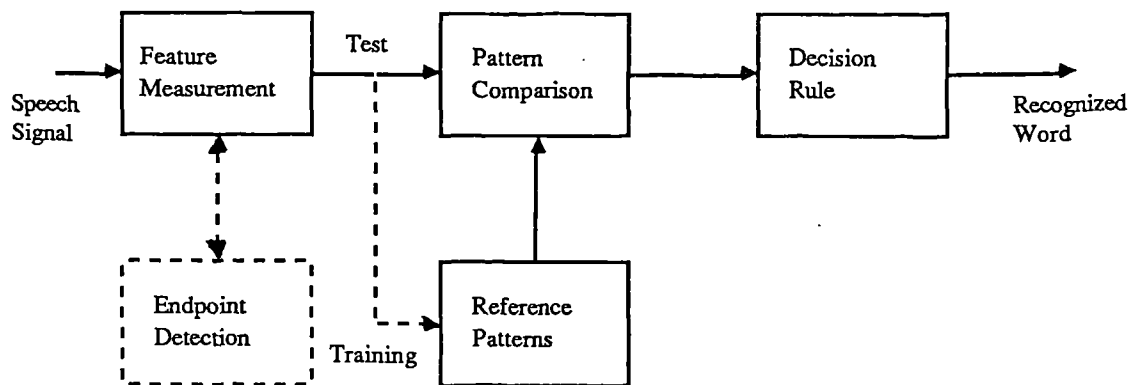


Fig. 2.2 . Pattern recognition model

The endpoint detection stage, which locates the beginning and end points of a word, can be included in the first stage. As we saw in the previous section, a short interval (frame) of speech is represented by an N-dimensional feature vector (or as a point in an N-dimensional feature space), and an utterance is represented by a sequence of vectors.

These vectors define a pattern in an N-dimensional space. During the training phase, a pattern is created for each reference word in the vocabulary. During the test phase, similarity measurements are made to compare an unknown pattern with each reference pattern. Then, a decision rule is used to choose the reference pattern which best matches the unknown pattern (or word). Before going into details, a few comments can be made regarding endpoints detection.

a. Endpoint Detection

Endpoint detection means finding the spoken word in the designated recognition interval: in other words, separating the speech signal from background sounds or noise. Different methods have been proposed for locating the first and last points of a word [43, 44, 45] where some feature measures like energy and zero-crossing rate have been used.

The endpoint detection is crucial in the recognition of isolated words for two reasons:

- Error in the endpoints location increases the probability of making recognition errors.
- Proper location of the endpoints keeps the overall computation of the system to a minimum.

The complexity of the word boundary detection depends on the speaking environment (e.g., speaking in sound proof-booth, or in computer room, or via telephone line, etc.), and on the transducer (e.g., telephone handset, high quality microphone, noise-cancelling microphone, etc.).

Endpoints detection is a very simple procedure when using a close-talking noise-cancelling head-mounted microphone, but becomes very difficult when the recording conditions degrade, especially when a word starts or ends with a weak fricative [46].

Apart from environment noise, the noise most harmful to the recogniser is often generated by the speaker himself through breath noise, such as aspiration or exhalation after speaking a word and quick inhalation or lip pops immediately before speaking. Inhaling produces no significant direct air blast on the close-talking microphone, whereas exhaling can produce signal levels comparable to speech levels.

b. Pattern Similarity Measures

In the recognition system a comparison is carried out to determine the similarity measure between a test (unknown) pattern or template, and all reference templates. One major difficulty in this case is that the speech utterances are rarely of equal temporal length. Their durations are dependent on the speaking rate, where we may have different duration for the repetition of the same word by the same speaker and across speakers. Therefore pattern similarity involves both time alignment (time warping) and distance computation, where these are often achieved simultaneously.

i. Dynamic Time Warping

Time alignment means the process of non-linear warping of a template in an attempt to align (synchronise) similar acoustic segments in the test and reference templates. This procedure, called dynamic time warping (DTW), combines alignment and distance computation through a dynamic programming procedure [47, 48]. Normalisation by means of time warping is an exceptionally powerful device and has contributed greatly to the accuracy of recognition systems.

Figure 2.3 shows an example of nonlinear time alignment of a test pattern $T(n)$, which has N frames or vectors, and a reference pattern $R(m)$, which has M frames.

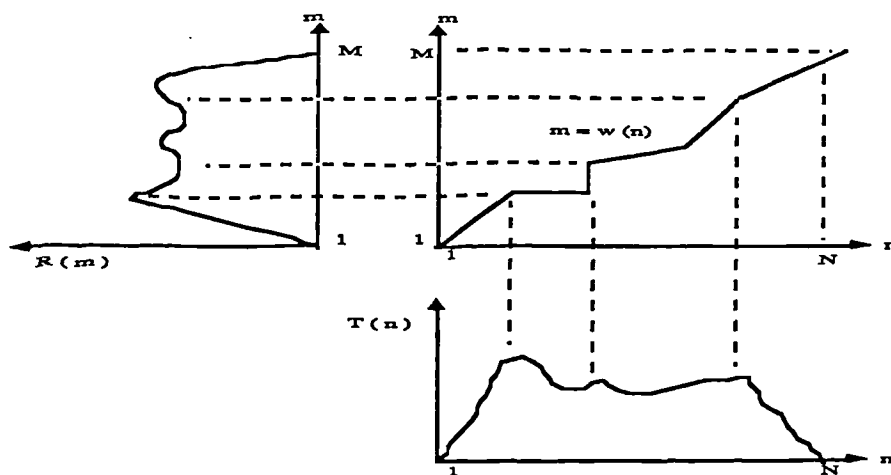


Fig. 2.3 Example of DTW of a test and reference patterns

DTW finds a warping function (or path) $m = w(n)$, which maps the time axis n of the test reference onto the time axis m of the reference pattern. Going frame by frame (or vector by vector) through the test pattern, DTW searches for the best frame in the reference pattern against which to compare each test frame. The warping curve is determined as the solution to the optimisation problem:

$$D = \min_{w(n)} \left[\sum_{n=1}^N d(T(n), R(w(n))) \right] \quad (2.9)$$

where $d(T(n), R(w(n)))$ is the distance between frame n of the test pattern and frame m of the reference pattern (which will be explained in the next paragraph). D is the minimum distance measure corresponding to the best path $w(n)$ through a grid of $N \times M$ points.

Restrictions on the time warping function have been studied by many researchers [49, 50, 51]. The restriction is achieved by reducing the search area of the dynamic programming and consequently the number of distances to be computed. Some of these restrictions are :

- Endpoints constraints on the path.
- Local path continuity constraints (i.e., the possible types of motion such as directions and slopes of the path).
- Global path constraints (i. e., the limitation on where the path can fall in the (n, m) plane.
- Distance measure (i. e., the type of employed distance).

Figure 2.4 displays some global and local continuity constraints imposed on the DTW path.

It is often difficult to locate word boundaries consistently, especially for isolated words spoken in a background of noise. Special DTW procedures have been used to eliminate the first few or last few frames from the total distance measure, by relaxing the test axis and/or the reference axis at the endpoints and relaxing the local continuity constraints at the endpoints [52].

DTW has recently been applied to the problem of training recognisers via automatic gathering of statistics on natural speech. Extracting training data or templates for short acoustic segments from continuous or connected speech requires tedious hand segmentation and labelling. By relaxing local continuity constraints, DTW has been used to align phones (e.g., phonemes, diphones, etc.) in unlabelled natural utterances with both synthetic and previously labelled natural utterances [53, 54]. Labelled utterances permit automatic extraction of sub-word templates from continuous speech.

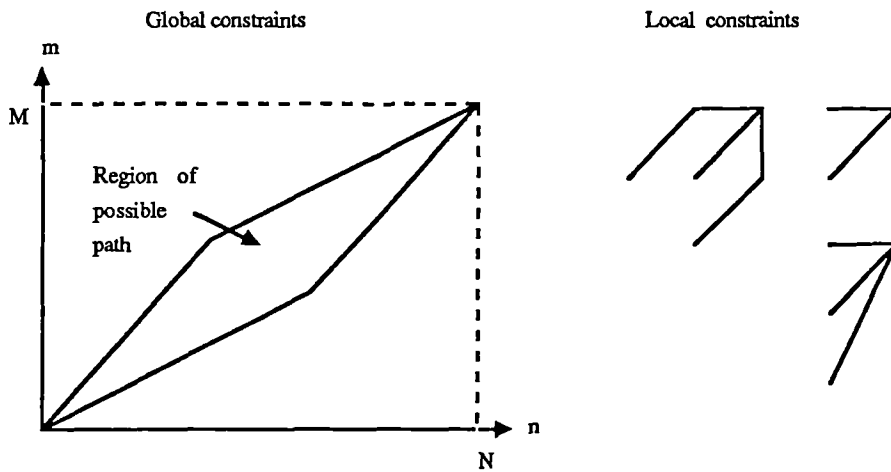


Fig. 2.4 Global and local continuity constraints on the DTW path

ii. Distance Measure

In order to implement the optimisation of Eq. (2.9) (and also for other purposes), a concept of distance (or distortion) between frames must be defined. Several possible distance measures can be used, depending on the type of the feature set [55].

Mahalanobis Distance

This distance, also called the covariance-weighted distance, is defined as:

$$d(T,R) = (T - R)^t W^{-1} (T - R) \quad (2.10)$$

where $d(T,R)$ is the distance between test and reference Frames. T and R are the N -dimensional feature vectors representing test and reference frames respectively. W^{-1} is the inverse of the autocorrelation matrix of the reference feature vector, and $(T - R)^t$ is the transpose of the vector $(T - R)$. Despite the theoretical advantages of the Mahalanobis distance, most recognisers use the Euclidian distance or the LPC distance described below, because it is difficult to reliably estimate W^{-1} from limited training data, and the latter two distances require less computation.

Euclidian Distance

The Euclidian distance is a result of setting the covariance matrix W^{-1} in Eq. (2.10) to be the identity matrix I . Then the Euclidian distance becomes:

$$d_E(T,R) = \sum_{n=1}^N (T(n) - R(n))^2 \quad (2.11)$$

where $T(n)$ and $R(n)$ are the n th components of the vectors T and R , respectively.

City-Block Distance

The City-block distance is defined as:

$$d_{CB}(T,R) = \sum_{n=1}^N |T(n) - R(n)| \quad (2.12)$$

It was found that there is no significant difference in the performance of an isolated word recognition system using the City-block distance instead of the Euclidian distance [59].

LPC Distance

For feature sets based on LPC parameters, an efficient distance measure was proposed by Itakura [47], and is called the log-likelihood ratio, or simply the LPC distance:

$$d_{LP}(T,R) = \log \left(\frac{a_R \quad V_T \quad a_R^t}{a_T \quad V_T \quad a_T^t} \right) \quad (2.13)$$

where a_T and a_R are the LPC feature vector of the reference and test frames, respectively, and V_T is the matrix of autocorrelation coefficients of the test frame. This distance is related to the spectral differences between two LPC modes. A different form of the LPC distance, which simplifies the computation of Eq. (2.13), is as follows:

$$d_{LP}(T,R) = \log \left(\sum_{n=1}^N V'_T(n) \quad R'_R(n) \right) \quad (2.14)$$

where N is the degree of the LPC model (or the dimension of the feature vector), $V'_T(n)$ is the autocorrelation vector of the test frame normalised by the LPC error, and $R'_R(n)$ is the autocorrelation of the reference vector.

c. Decision Rules for Recognition

The last major step in the pattern recognition model of Figure 2.2 is the decision rule, which chooses the reference pattern (or patterns) that matches most closely the unknown test pattern. Most isolated word recognition systems use the nearest neighbour rule (NN), or the K-nearest neighbour rule (KNN). The NN rule chooses the pattern R^{i^*} with smallest average distance as the recognised pattern, according to the following equation:

$$i^* = \underset{i}{\operatorname{argmin}} [D^i] \quad (2.15)$$

In some systems, where there are several stored reference templates for each vocabulary word (corresponding to pronunciations of the word by several speakers, or several repetitions of the word by one speakers), The KNN rule may be applied. This rule finds the nearest K neighbours (among all templates) to the unknown and chooses the word with the maximum number of entries among the K best matches [57]. In an alternative modified KNN rule, the selected output corresponds to the word that minimises the average distance between the test template and the best K matches for each vocabulary word. The KNN rule (with K of 2 or 3) improves the recognition accuracy when the

number of templates per word is above 6, but increases the necessary computational requirement [58].

d. Creating Reference Templates

In the recognition system of Figure 2.2, a training phase is assumed before an actual recognition can take place. The simplest speaker-dependent systems employ causal training, in which each speaker utters every word in the vocabulary one or more times, and a reference template is created. Since speakers tend to pronounce a given word differently at different times or in different contexts (because of different articulatory structure for different speakers), a few repetitions of each word are often used in training. Most speaker-dependent systems use 1-3 templates per word, while speaker-independent systems (multi-speaker systems) use 10-12 [59].

Reducing the number of templates for each reference word to a reasonable number (as mentioned above) is necessary to reduce confusions and storage requirements in speech recognition systems. Two methods are used for creating reference templates, namely averaging and clustering.

In averaging, all the occurrences of a given word are averaged together, after some form of time alignment. This gives a single reference template for a speaker-dependent system [60]. For a speaker-independent system, averaging can create an unrepresentative pattern if the templates differ substantially.

In a speaker-independent system, at least 100 speakers must provide multiple training tokens for each word, which implies that a substantial clustering is necessary to merge the tokens to a representative set of 10-12 templates for efficiency. The K-means clustering method [61], and the unsupervised K-means clustering without averaging method [62], have been used. In clustering, the N templates of each vocabulary word are grouped together to form M clusters, using the nearest neighbour rule. For each such cluster, a single template is created using averaging technique over the tokens of that cluster.

d. Results of Pattern Matching Approaches in Isolated Word Systems

The accuracy of isolated word recognition systems using pattern matching techniques

varies from 90-100% [63, 64], according to the following factors:

- Vocabulary size
- Vocabulary complexity
- Speaker-independent or dependent
- Number of templates per word

The accuracy is actually a function of vocabulary complexity as much as or more than a function of vocabulary size, especially when the vocabulary includes many similar sounding words (i.e., acoustically similar words), such as the alphabet words (e.g., B, D, E, G, P, T, V). The poor performance of acoustically similar words can be improved by introducing a two-pass recognition method [65], where the first pass decides the equivalent class of the unknown word, and the second pass looks for the equivalent word within the specified class. A normal distance measure is used in the first pass and a weighted distance is used in the second pass, which would help discriminating between acoustically similar words. The two-pass approach has also been used in a large vocabulary system [66]. In the initial pass, linear matching between the test word and the reference using a few features (e. g., duration and two or three average spectra) can be used. The aim of this initial pass is to reduce the number of candidates to be considered in the detailed second pass which uses DTW.

e. Advantages and Disadvantages of the Pattern Matching Approach

i. Advantages

- The pattern or template matching model of Figure 2.2 can be used with any word vocabulary.
- It can be used as either a speaker-dependent or a speaker-independent system.
- It is modular in its three main stages and alternative algorithms (i.e., new feature sets, new DTW methods, etc.), can be readily employed and tested.
- It uses no speech specific knowledge.

ii. Disadvantages

- A large amount of storage is needed for storing reference templates.
- Heavy computation is required for time alignment between test and reference

patterns (i.e., DTW and distance measures).

- The amounts of required storage and computation increase linearly with the vocabulary size.

f. Other Techniques Employed in Template Matching Model

Several other techniques have been introduced to overcome the problem of high storage and computational requirements without significant degradation in recognition performance such as:

i. Trace Segmentation Method

In this method the number of frames of each utterance is reduced by exploiting the stationarity of speech segments [56]. An utterance can be seen as a sequence of points (trace) in an N-dimensional feature space. The stationary parts of the speech signal cause a high point density along this trace, while the transitional parts (rapid spectral changes) with short duration lead to points that are spaced far apart on this trace.

The idea of trace segmentation method is to represent a word by a fixed number of points (segments) uniformly spaced along the trace, thereby replacing a time domain sampling with a sampling in the N-dimensional space.

In this method two positive effects are obtained, namely a reduction in the number of frames and better allocation of points along the trace (eliminating some redundancy). Experiments showed that 99% accuracy had been achieved for a speaker-dependent system of 31 words, using a 13-channel filter bank [56].

ii. Transient Matching Method

In this method, only the transient parts of the speech are used as recognition elements to achieve reduction in storage and computation [67].

iii. Vector Quantisation Method (VQ)

In this method, a substantial cut in the storage and computation can be achieved. Each vector in the speech template can be replaced by the address of the closest codeword in

the codebook. The codebook is designed by minimising the average distance between the designed codebook vectors and a large number of appropriate feature vectors. A codeword represents an entry in the codebook.

If only the reference templates are quantised, the method is called single-split VQ, but if both reference and test templates are quantised, the method is called double-split VQ. In double-split VQ method, a pre-computation of the distance matrix of each codeword vector to every other codeword vector reduces the problem of distance computation in the time alignment algorithm to a simple table-lookup operation.

The performance degradation, due to the distortion introduced by vector quantisation, was found small [68].

2.3.3 Recognition Without Time Alignments

Several methods were proposed to avoid the process of time alignment (DTW), and the most important one is the use of a vector quantisation technique. In this case, a separate VQ codebook was designed for each word in the vocabulary by using data containing several repetitions of each word. An unknown test word is classified by quantising each frame in the word using each available codebook (reference word). The average distortion over all frames of the input word is then computed for each codebook and the input is classified as the word corresponding to the codebook yielding the lowest average distortion. The accuracy of such a recognition method was 99% for a small set of words (20 words) [69], using codebooks of 32 and 64 entries for a speaker-dependent system. A speaker-independent experiment using 9 utterances per word, 9 different codebooks, gave an 87% accuracy by employing codebooks of 128 entries.

The performance of such methods was improved by incorporating some time sequence information [70, 71]. In this case, words in the training and input sequences were normalised linearly to the same length, and then divided into sections. A separate codebook was then designed for each section of each vocabulary word. Each word was thus represented by a time-dependent sequence of section-codebooks. New words were classified by performing VQ and finding the multi-section codebook that achieves the smallest average distortion. Results on 20 words for a speaker-independent test gave 97% recognition accuracy with small size codebooks (to achieve faster computation).

2.3.4 Hidden Markov Modelling (HMM)

A probabilistic function of hidden Markov chain is a stochastic process generated by two inter-related mechanisms: an underlying Markov chain having a finite number of states, and a set of random functions, one of which is associated with each state. At a given time instance, the hidden Markov process is in a unique state and an observation is generated by the random function associated with the state. This causes the underlying Markov chain to change state in accordance with its transition probabilities. These states can not be observed directly (hidden), but only the outputs of the random functions at each state are seen.

It is quite reasonable to consider the speech signal as being generated by such process. We can imagine the vocal tract as being in one of a finite number of articulatory configurations or states. In each state, a short (in time) signal is produced that has one of a finite number of prototypical spectra depending on the state. Thus, the power spectra of short intervals of the speech signal are determined solely by the current state of the model, while the variation of the spectral composition of the signal with time is governed predominantly by the probabilistic state transition law of the underlying Markov chain.

In principle, the Markov chain may be of any order, and the outputs from its states may be multivariate random processes having some continuous joint probability density function. The most common network for automatic speech recognition is the first-order Markov processes, i.e., for which the probability of transition to any state depends only upon that state and its predecessor. Higher-order Markov processes could exploit restrictions on which sounds may occur in sequence within words, but the computational complexity of such models has thus far precluded their application to acoustic analysis in automatic speech recognition.

Speech recognition uses processes whose observations are drawn from a discrete finite alphabet according to discrete probability distribution functions associated with the states. Let us assume that the underlying Markov chain has N states:

$$q_1, q_2, \dots, q_N$$

and the observations are drawn from an alphabet V , of M prototypical spectra

(codebook of M codewords):

$$v_1, v_2, \dots, v_M$$

The underlying Markov chain can then be specified in terms of an initial state probability vector:

$$\pi = (\pi_1, \pi_2, \dots, \pi_N)$$

and a state transition matrix:

$$A = \begin{bmatrix} a_{ij} \end{bmatrix} \quad 1 \leq i, j \leq N$$

Here π_i is defined as the probability of q_i at some arbitrary time, $t = 0$, and a_{ij} is the probability of transiting to state q_j given the current state, q_i , that is:

$$a_{ij} = \text{prob}(q_j \text{ at } t+1 \mid q_i \text{ at } t)$$

The random processes associated with the states can be collectively represented by another stochastic matrix:

$$B = \begin{bmatrix} b_{jk} \end{bmatrix} \quad 1 \leq j \leq N \quad \text{and} \quad 1 \leq k \leq M$$

where b_{jk} is the probability of observing symbol v_k given current state q_j :

$$b_{jk} = \text{prob}(v_k \text{ at } t \mid q_j \text{ at } t)$$

This hidden Markov model M , is identified with the parameter set (π, A, B) .

In order to use hidden Markov modelling, two specific problems must be solved to perform speech recognition:

- Observation sequence probability estimation, which will be used for classification of an utterance.
- Model parameter estimation, which will serve as a procedure for training models for each vocabulary word.

Both problems proceed from a sequence O , of observations:

$$O_1, O_2, \dots, O_T$$

where each O_t for $1 \leq t \leq T$ is some $v_k \in V$.

An efficient method for estimating parameter sets of hidden Markov models is given in reference [72]. Figure 2.5 shows a five-state model [73]. In this left-to-right model, it always begins in state q_1 (i.e., the initial state probability is $\pi_1 = 1$, and $\pi_i = 0$ for $i \neq 1$), and ends with state q_5 without revisiting states which have been left.

In this model, three transitions are allowed from each state:

- A loop transition back to the same state (representing the insertion of an acoustic segment).
- A transition to the next state (a substitution or a new segment).
- A skipping transition to the following state (corresponding to the deletion of the acoustic segment of the skipped state).

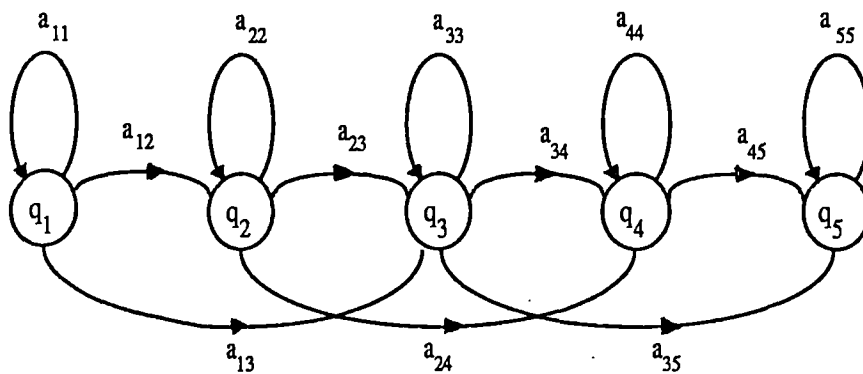


Fig 2.5 Five-state left-to-right Markov model

More general HMMs, allowing transition to any succeeding state (rather than limiting the transition from state i to only states $i, i+1, i+2$) increase computation and do not improve the recognition performance [73].

Isolated word recognition using HMM consists of two phases, training and recognition. In the training phase, the training set of observations (feature vectors) is used to derive a set of reference models, one for each word in the vocabulary. In the classification phase, the probability of generating the test observation is computed for each reference model. The test is classified as the word whose model gives the highest probability.

In the training phase, the alphabet V , or the set of spectral shapes, is generated by performing a vector quantisation on a large number of training feature vectors, where an N entries codebook can be designed. Then, the feature vectors of all training words are described by indices of this codebook. A Markov model for each word is generated from a large number of repetitions of this word, resulting in a number of models equal to the number of vocabulary words used in a particular recognition system.

In the recognition phase, assume a set of R words (W_1, W_2, \dots, W_R) represented by R models (M_1, M_2, \dots, M_R), an unknown word is represented by an observation sequence:

$$O = O_1 \ O_2 \ \dots \ O_T \quad O_t \in V, \text{ and } 1 \leq t \leq T$$

The probability of O having been generated by model M_r is :

$$P_r(O/M_r) = \sum_{i_1, i_2, \dots, i_T} \pi_{i_1} \ b_{i_1}(O_1) \ a_{i_1 i_2} \ \dots \ b_{i_T}(O_T) \ a_{i_{T-1} i_T}$$

The unknown utterance is classified as W_r if, and only if $P_r \geq P_j$ where $1 \leq j \leq R$ and $r=1, 2, \dots, R$. The interpretation of the computation in the above equation is the following. Initially (at time $t-1$) we are in state i_1 , with probability π_{i_1} , and generate the symbol O_1 with probability $b_{i_1}(O_1)$. We then make a transition to state i_2 with probability $a_{i_1 i_2}$, and generate symbol O_2 with probability $b_{i_2}(O_2)$. This process continues until we make the last transition for state i_{T-1} to state i_T with probability

$a_{i_{T-1} i_T}$ and generate symbol O_T with probability $b_{iT}(O_T)$, where i_1, i_2, \dots, i_T form the optimal state sequence related to the observation sequence $O_1 O_2 \dots O_T$. A detailed and computationally efficient algorithm for evaluating the above equation is given in [75, 76].

The training procedure for such a system is computationally expensive, but it needs to be done only once. In the classification mode, a hidden Markov model with vector quantisation requires 17 times less computation than the classical method using DTW, and also 10 times less storage is needed. In a speaker-independent system using the 10 digits as vocabulary words, an average recognition accuracy of 96% was achieved [77].

Dynamic time warping can be considered as a special case of hidden Markov modelling [78]. In the pattern matching approach, words are represented by a sequence of feature vectors (frames), and the DTW looks for the optimal path between the test and reference frames (Viterbi algorithm). If each frame is considered as a hidden Markov state, then the DTW path is equivalent to the most likely state sequence (e.g., a Viterbi state sequence for an observation sequence of length T , and N state hidden Markov model). HMM requires much computation during the training phase when the model is built, but much less so during classification, provided the number of states is much less than the number of frames in the speech utterance.

2.3.5 Neural Networks

In recent years, the advent of new learning procedures and the availability of high speed parallel supercomputers, have given rise to a renewed interest in parallel distributed processing models known as artificial neural networks or simply neural nets. These models attempt to achieve good performance via dense interconnection of simple computational elements. The neural nets are particularly interesting for cognitive tasks that require massive constraint satisfaction, i. e., the parallel evaluation of many clues and facts, and their interpretation in the light of numerous interrelated constraints. Cognitive tasks, such as vision, speech, and language processing, are also characterised by a high degree of uncertainty and variability and it has proven difficult to achieve good performance for these tasks using standard sequential programming methods. In general, such constraints are too complex to be easily programmed and require the use of automatic learning strategies, which are now available [78]. Learning or adaptation is a major focus of neural nets research. The ability to adapt and continue learning is essential

in areas such as speech recognition, where training data is limited and new talkers, new words, new dialects, new phrases, and new environments are continuously encountered.

In pattern recognition systems, the major problems are the time axis distortion and spectral pattern variation. The former problem has been mathematically well modelled and solved by the use of DTW. On the other hand, the spectral variation, which is caused by a complex mixture of several effects, is hard to treat. The neural net is quite a general pattern recognition model which, by being fed training samples of given categories, can learn to achieve a function to discriminate between the categories. Therefore, it is suitably applicable to pattern recognition problems where an analytical approach is inapplicable. This, in turn, implies the usefulness of the neural model in solving spectral pattern variation problems.

Experiments on using the neural nets for speaker-independent recognition gave 95% for 20 isolated words [79], and 98% accuracy for 10 isolated words [80], using different neural nets implementations. These results suggest that appropriately designed artificial neural networks are well-suited for a speaker-independent recognition task.

2.3.6 Acoustic-Phonetic Approaches

The whole word pattern-matching methods described in the previous sections are usually used with vocabulary sizes ranging from a dozen to a few hundred words. These methods have some limitation concerning the vocabulary size, which are in brief:

- The amount of required storage and computation becomes excessive. Even with present technology, this limitation is still important but is becoming less serious with each year of technological developments.
- The time needed for the enrolment process of new speakers (uttering all words, or generate stochastic models, etc.), would limit the use of these systems in several applications.
- When the number of words is large (particularly for continuous speech), variations in the pronunciation of one word will often exceed the measured differences between repetitions of different words. Under these circumstances recognition errors will be unavoidable (where a small irrelevant difference in articulation of a phoneme may give a greater accumulated matching error).

For all the above reasons, most large vocabulary speech recognition systems are using acoustic-phonetic approaches.

In acoustic-phonetic approaches, the recognition units are smaller than words (sub-word), where some form of signal to symbol transformation is performed. These units are either phonetic segments or acoustic homogeneity segments, such as phonemes, diphones, demisyllables, syllables, or crude phonetic segments. Employing these units as recognition units facilitates the utilisation of linguistic information (or knowledge) to manipulate the results of the acoustic-phonetic classification process in speech recognition systems.

Figure 2.6 shows a simplified block diagram for a speech recognition model based on the acoustic-phonetic approaches. In this model, the speech signal is divided into segments according to specific acoustic-phonetic and phonological rules, and a labelling scheme associates a phonetic symbol with each segmental unit. The choice of acoustic features in the feature measurements stage is influenced by the segmentation strategy. These could vary from parameters related to a speech production model, to parameters related to the auditory physiology and psychophysics, as was demonstrated in the feature measurement section (Section 2.3.1).

In this model, each vocabulary word is represented by a string of segmental labels and stored in a lexicon. For an unknown word, there exists one (or more) word candidate. At the last stage of the model, some decision rules are used to locate the most likely word candidate.

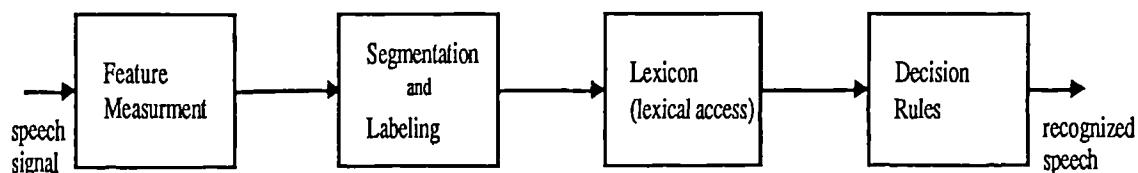


Fig. 2.6 General model for acoustic-phonetic approaches

The model of Figure 2.6 can be extended to continuous speech recognition by incorporating some knowledge sources, as we will see in the following sections. For isolated words, it requires some acoustic-phonetic rules and phonological rules to facilitate and enforce the recognition process where these rules depend on the type of the recognition unit. However, the first step in a recognition system of this kind is to decide which units are to be used. In the following paragraphs some of these recognition units are described.

a. Phonemes

Different sets of phonemes are used in different languages, their number varies between 40-60 phonemes for English, Arabic and most of the other European languages. However, phonemes have a number of contextual variations known as allophones and there are some 100-200 of them, according to the specific language. The problem with phoneme recognition units is segmentation, where generally, the acoustic manifestation of each phoneme is modified by co-articulation effects. The successive phonemes influence one another or even overlap. For this reason, phoneme segmentation is rather difficult and inaccurate.

b. Diphones

A diphone (phone-pair) is defined as the interval from the middle of one phonetic segment to the middle of the next. The transitions between adjacent phonemes are included in the diphones. There are about 1200 to 1500 different diphones in English [81], and about the same number in Arabic [4].

c. Syllables and Demisyllables

Coarticulation occurs mainly within a syllable, and also across syllable boundaries. This has led to the idea, that preliminary segmentation of the speech signal should be syllabic rather than phonemic. The syllable can be considered as an articulatory as well as perceptual processing unit. It has been estimated that the number of syllables in English is about 10,000 [82].

The number of syllabic units in a language can be considerably reduced by dividing each syllable into two parts, one demisyllable containing the initial part of the syllable, and one

demisyllable containing the final part of the syllable. The cutting point can lie within the syllable nucleus (at the point of maximum intensity). The resulting inventories of initial and final demisyllables contain in the order of 2000 elements for English [83], and the number of demisyllables in other languages may be similar to this number [82].

A drastic reduction in the number of units is reached by dividing the syllable into three parts: the syllable nucleus, the initial consonant cluster preceding the nucleus, and the final consonant cluster following the nucleus. It was found, that 100 important demisyllables out of the 200 available demisyllables (consisting of 47 initial consonant clusters, 153 final consonant clusters which include vowels), are enough to describe the majority of syllables for German language [84]. English has about 70 initial consonant clusters, about 100 final ones, 12 vowels, and 9 diphthongs [82].

d. Acoustic Sub-Word Units

The major problem with using the above sub-word speech units is, that robust and reliable algorithms for automatically determining the presence and/or identity of such units do not yet exist. Those sub-word units have been defined based on a linguistic description of the language. The acoustic sub-word units (ASUs) are derived acoustically without any reference to linguistic content. A small set of ASUs can be created acoustically from the speech signal over a wide range of training speech data [85, 86].

The well-defined linguistic sub-word units make lexical decoding an easy task, since a standard dictionary of pronunciation will generally provide a simple and straightforward mapping between the chosen linguistic units and the word orthography. In contrast to that, ASUs have no simple linguistic interpretation, and lead to great difficulties in lexical decoding, since no simple and/or straightforward mapping to words is possible.

e. Broad Phonetic Classes (units)

Phonemes are the smallest set of linguistic units, but unfortunately, it is often difficult, if not impossible (so far) to identify phonemes reliably from the acoustic speech signal. Instead of that, phonemes are divided into sub-groups each of which contains a number of phonemes sharing almost the same acoustic properties. These sub-groups are called broad phonetic classes, e.g., vowels, plosives, fricatives, nasals, liquids, and semivowels. These classes are related to the manner of articulation (see chapters 3 and 4

for more details). In this case, a crude, but reliable, acoustic analysis is performed in terms of broad phonetic classes [87].

Words in the lexicon are described according to their broad phonetic labels. Unlike the use of acoustic sub-word units, linguistic knowledge is still appropriate to these representations (with some inevitable modification).

An acoustic-phonetic approach, which uses a hybrid scheme of broad phonetic classification for consonants and detailed vowel recognition, is investigated in Chapter 4.

In general, the process of segmentation and labelling is very error prone. The degree of error depends on the actual recognition units. In the decision rules stage of Figure 2.6, different strategies can be used to correct most of the errors made at the segmentation level, through the use of different linguistic sources of knowledge.

2.3.7 Syntactic Pattern Recognition

Syntactic pattern recognition has been applied to connected and continuous speech recognition [93]. It has also been applied to isolated word recognition [88]. Figure 2.7 shows a block diagram for such a system.

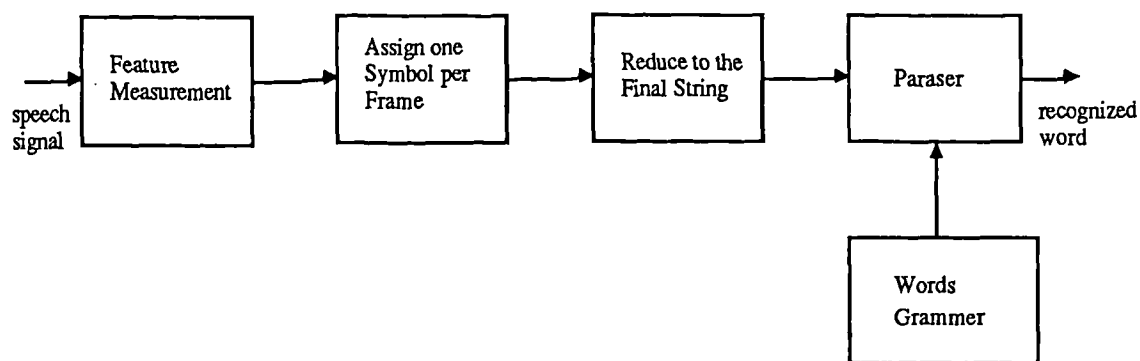


Fig. 2.7 Syntactic pattern recognition

In this method, each feature vector is assigned one symbol, and then a reduction process is carried out to result in a string of symbols or primitives, which is a short characteristic

representation of the utterance (represents the acoustically consistent region along a word). Syntactic pattern analysis is then performed on the symbol string using a data base of context-free grammars representing the word in the vocabulary.

Constructing a database of grammars for the vocabulary words is performed using symbol strings generated from training word utterances, representing the set of vocabulary words uttered by several speakers. This database of grammars contains one context-free grammar for each vocabulary word. In the parser stage, when the final input is accepted by a particular grammar, the input utterance is recognised to be the vocabulary word corresponding to that grammar. Recognition accuracy of such systems is very high for a small-vocabulary speaker-independent system [89].

2.4 Connected Word Recognition

In connected word recognition, the spoken input is a sequence of words from a specified vocabulary, and the recognition is based on isolated word recognition. This is in contrast with continuous speech recognition, which generally involves recognition from linguistic units. Typical examples include connected digit string, where the vocabulary is the set of 10 digits (0-9), or connected letter recognition (e.g., for spelling words, names, etc.), where the vocabulary is the set of the alphabet. A pattern recognition model had been used for connected word recognition, as in Figure 2.8.

In this model, the patterns to be matched consist of a test utterance (a string of words), containing a set of word templates. For a small number of words in the test utterance and vocabulary (such as 2 or 3), it is possible to concatenate in all possible orders to form a set of connected word templates, then to apply dynamic programming to determine the sequence of templates which best match the test utterance. For large strings and bigger vocabularies, the amount of computation involved with this approach rapidly becomes prohibitive. Syntactic constraints on the appearance and order of words for a given application may reduce the number of comparisons.

Several methods have been proposed to reduce the calculation of DTW. One is two-level dynamic programming, which compares templates in two levels, one for individual words and the other for the entire phrase [90].

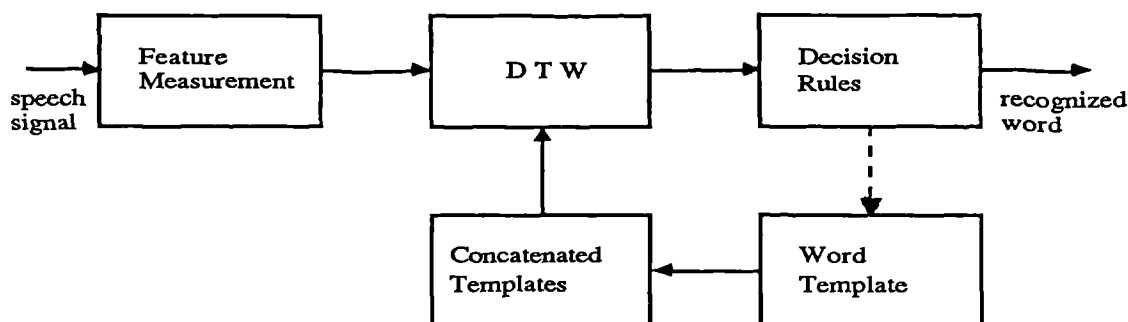


Fig. 2.8 Connected word recognition model

A more efficient algorithm, called 'level-building dynamic time warping', has been proposed [91]. In this algorithm, matching is performed level by level, where each word reference template R_i represents a level. At the first level, DTW is applied to compare R_i for each word that may appear in the initial position of a test utterance against the initial position of the test template T . For each comparison, distance scores are stored for all allowable endpoints in T , subject to the normal continuity constraints. At level 2, R_i for all possible second position words are compared against T , with paths starting from the endpoints of the previous level and proceeding to allowed endpoints for second word. This procedure continues until all levels have been processed. Figure 2.9 displays possible paths through a four-level phrase with a parallelogram warping window. The endpoints e_i of the words in the connected phrase are determined by backtracking.

Another method related to the level-building method is called the one-stage approach [92]. This method requires much less warping memory, and for long utterances significantly less computation, than the level building method. Each reference template is matched against the first part of the test utterance (phrase) and the optimum path, yielding the minimum distance score, is determined similarly as in isolated word recognition. Each reference template is then matched against the test utterance starting at the point where the last template match ended. Computation proceeds for all templates in parallel, in one pass through the test utterance. This process continues until the end of the test utterance is reached. At this point the test score corresponding to the word at the end of the test utterance is selected. The sequence of reference templates which leads to this score is chosen as the string of words in the input pattern.

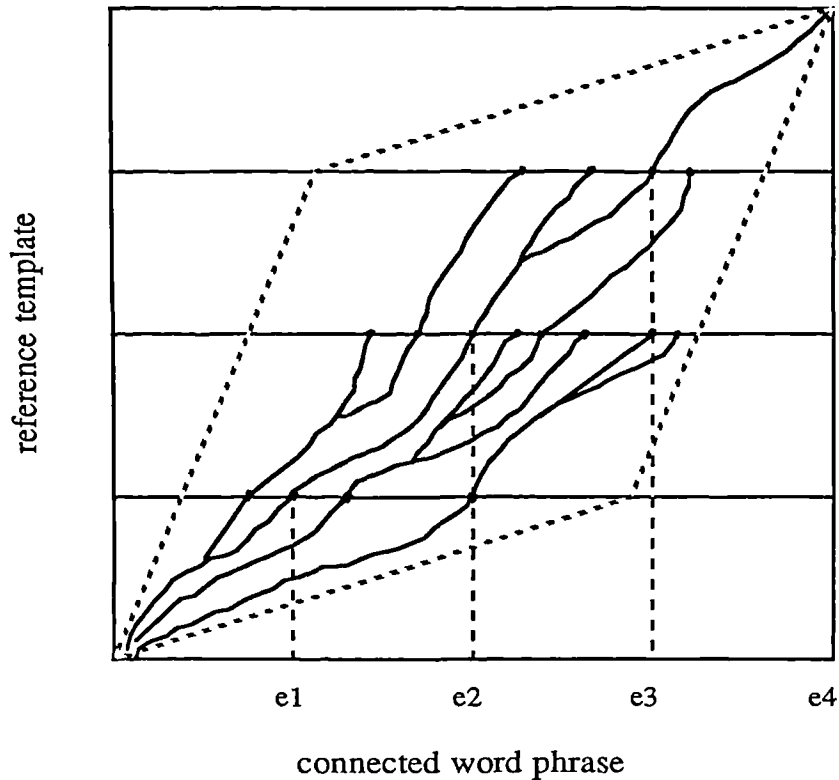


Fig. 2.9 Possible paths through a four-level phrase with a parallelogram warping window

It was shown that the one-stage algorithm requires only about 4% of the computation required by the two-level algorithm, and about 25% of that of the level-building algorithm [92]. In addition, the one-stage algorithm uses only about 10% of the storage requirements of the other above mentioned algorithm, during DTW computation.

One of the advantages of both the one-stage algorithm and the two-level algorithm is that syntactic constraints may easily be incorporated. The point in the test utterance where one template changes to another is determined as a part of the optimisation process. If certain word orders are unlikely, or prohibited, appropriate weights can be inserted in the distance computation.

2.5 Continuous Speech Recognition

In the production of continuous speech, the variation of the acoustic realisation of the speech is more acute compared to the case of producing isolated words. In continuous speech, pronunciation is less careful, the speaking rate is less constant, speaker differences are underlined, coarticulation effects exist between words as well as within them, and acoustic segments are commonly deleted, inserted, or substituted. Although speech is perceived as a sequence of separate words, there is often little evidence of word boundaries. In addition to that, the importance of a word in the message affects its stress and intonation and hence its acoustical realisation.

In attempting to recognise continuous speech, the machine is presented with an input message, which may be imprecise, and in which not all of the information necessary for decoding is unambiguously or completely encoded. The listener uses his knowledge of the language in order to decode the speech message. The spoken message will not be understood unless the speaker and the listener use the same language. This means that human listeners make use of linguistic cues and constraints in recognising continuous speech. This leads to the importance of using sources of linguistic knowledge in automatic continuous speech recognition. These sources are :

- Acoustic-phonetic knowledge
- Phonological knowledge
- Prosodic knowledge
- Lexical knowledge
- Syntactic knowledge
- Semantic knowledge
- Pragmatic knowledge

Before going into some details about these sources, we should clarify the difference between continuous speech recognition and a speech understanding system.

2.5.1 Speech Understanding Concept

Continuous speech recognition attempts to transcribe speech onto orthographic form. The inherent variability of naturally spoken speech makes continuous speech recognition so difficult. Restrictions have been imposed on the continuous speech, where the utterances

to be recognised are spoken in a well defined context (related to a specific task), so that high-level knowledge sources such as syntax and semantics can be applied to support the basic recognition processes. These systems are called simply speech understanding, where the speech is first understood in order to be recognised through the incorporation of some high-level knowledge sources. The speech understanding systems are task-dependent systems. On the other hand, task-independent continuous speech recognition will allow a change in topics from one utterance to the next and still correctly identify the wording of the sentence.

2.5.2 Sources of Knowledge

A brief description of each source of knowledge is given below.

a. The Acoustic-Phonetic knowledge

To use acoustic-phonetic approaches in continuous speech recognition, it is necessary to take into account the large number of different acoustic cues related to particular minimal phonetic differences [84]. It is very difficult to formulate the rules for making some fine phonetic distinctions. A method commonly adopted for this process is to accept that some very skillful human beings (experts) can perform the interpretation task, through a spectrogram reading experiment [93]. Those rules can be transferred into machine, and stored into a knowledge base.

In a speech recognition system the acoustic-phonetic analyser extracts a set of acoustic parameters from the speech signal (see Section 2.3.1). Those parameters are segmented into linguistic units, and classified according to their structure, by exploiting specific acoustic-phonetic cues or rules from the acoustic knowledge source, perhaps using some form of 'expert systems'.

b. Phonological Knowledge

All languages are highly constrained, for example, by their phonetic inventory, which is only a small sample of all possible speech sounds. They are also constrained in the possible combination of these sounds through phonological rules. The presence of these constraints acts to reduce the amount of uncertainty in the phonetic string (or strings) which results from the acoustic-phonetic stage of a speech recogniser. On the other hand,

the co-articulation effect and the fluency of continuous speech affect the phonetic information in an utterance. Phonological rules express the systematic ways in which the realisation of words or phonemes (or other linguistic units) may change with their environment (and how these units deviate in 'defective' speech from the norm of the language) [94].

Generally speaking, the more we can use the phonological structure of the language and the more we can make use of this knowledge in a recognition system, the more successful we should be in our work in automatic speech recognition.

An appropriate set of acoustic-phonetic parameters can be extracted from the speech signal to provide a necessary and sufficient set of phonological features for further processing. Any desired lexicon can have its entries efficiently represented in the most useful phonological units, such as phonemes, allophones, diphones, demisyllables and syllables. None of these units is truly ideal for recognition, and for that reason, it may be well to consider the use of a combination of units in an automatic speech recognition.

c. Prosodic Knowledge

Prosodic or suprasegmental information, namely stress, intonation (variation of fundamental frequency F_0 with time [9]), pauses, and timing structures, can be extracted from the speech signal. This information offers an independent way of acoustically detecting some aspects of the syntactic structure of the speech, without depending upon the potentially errorful sequences of hypothesised words derived from the incoming acoustic-phonetic information. For example:

- Prosodic information can provide a variety of secondary aids to phonetic analysis, such as cues to voicing, location of syllable nuclei (vowels), guide for efficient acoustic analysis, etc.
- From pauses and very large F_0 variations, discourses can be divided into sentences, and sentences into clauses (boundary detection).
- Phrase categories can be determined from the aspects of stress patterns, general slope of F_0 contour (intonation), or other prosodic information.
- Important words in the sentence are more stressed than other words.

d. Lexical Knowledge

The lexicon, or vocabulary, allowed by each speech recognition or understanding system is represented internally in terms of pronunciations (phonetic transcription) of the words. Some systems encode all the multiple pronunciations of each word that arise from contextual effects.

e. Syntactic Knowledge

Syntactic knowledge enables the recogniser to determine whether a particular sequence of hypothesised words can occur within a grammatical sentence. In addition, syntax provides a basis for predicting additional but unhypothesised fragments of the sentence. The syntactic rules, or grammar, comprise a set of rules that specify legitimate linguistic expressions, and allowable combinations of words from the pre-defined vocabulary in the recognition system.

f. Semantic Knowledge

Semantic knowledge provides a capability to determine whether a syntactically correct sentence is actually meaningful or not. Semantic information is also employed to choose between words or sentences which seem equally likely on phonology, syntactic or other grounds.

g. Pragmatic Knowledge

Pragmatic knowledge enables the recognition system to determine whether a meaningful sentence is plausible and appropriate in the context of an ongoing dialogue. In a dialogue, a speaker's response must not only be a meaningful sentence but also a reasonable reply to what was said to him. The pragmatic knowledge can predictively constrain the types of sentences that might meaningfully prolong an ongoing dialogue.

Each type of the above higher-level knowledge defines additional constraints, where sentence interpretations must satisfy them. If properly exploited, these constraints can eliminate unlikely interpretations from consideration. As a consequence, These actions can reduce the number of incorrect hypotheses generated, extracted, or accepted.

2.5.3 Speech Understanding Models

Speech understanding systems attempt to integrate traditional speech recognition techniques with artificial intelligence (AI) techniques to give the extra power needed to deal with natural speech.

The actual mapping of an input utterance into a representation of its meaning takes place through a series of intermediate representations or levels. The translation from one level to the next higher level is accomplished by applying knowledge from one or more sources. Figure 2.10 shows one of the simplest arrangements for a speech understanding system, the so-called hierarchical structure [96]. In this model, the input speech is first analysed by the acoustic analyser to extract appropriate parameters, which provide the necessary features to describe the already chosen phonological units (let us assume phonemes). The acoustic analyser also provides sufficient parameters for prosodic feature extraction. The segmentation and labelling (into phonemes) are accomplished in the phonetic processor, through the use of acoustic parameters. The prosodic processor extracts prosodic features from the acoustic parameters and detects those prosodic cues related to the linguistic structure. The acoustic analyser, the phonetic processor, and the prosodic processor (only the prosodic feature extraction), can be combined into one stage as an acoustic-phonetic processor, sometimes called as the system front-end.

Because the acoustic-phonetic analysis is an inexact process, the output from the acoustic-phonetic processor is not a simple string of phonemes (or any other units), but a lattice of alternatives. This lattice is the first and lowest level of representation. The next higher level of representation is a network of word hypotheses which are grouped into syntactically legal phrase structures by the syntactic analyser to form the next level of representation. Finally the semantic and pragmatic processes are employed to determine the appropriate sentence which has the actual meaning.

In practice, the system may operate either in bottom-up mode, top-down mode, or a mixture of both. In bottom-up mode each level is derived directly from the level below. In top-down mode, each lower level is extended only when it is needed, in order to extend the higher level above. Bottom-up is effectively just another name for straight forward data-driven processing, where as top-down represents a hypothesis and test paradigm.

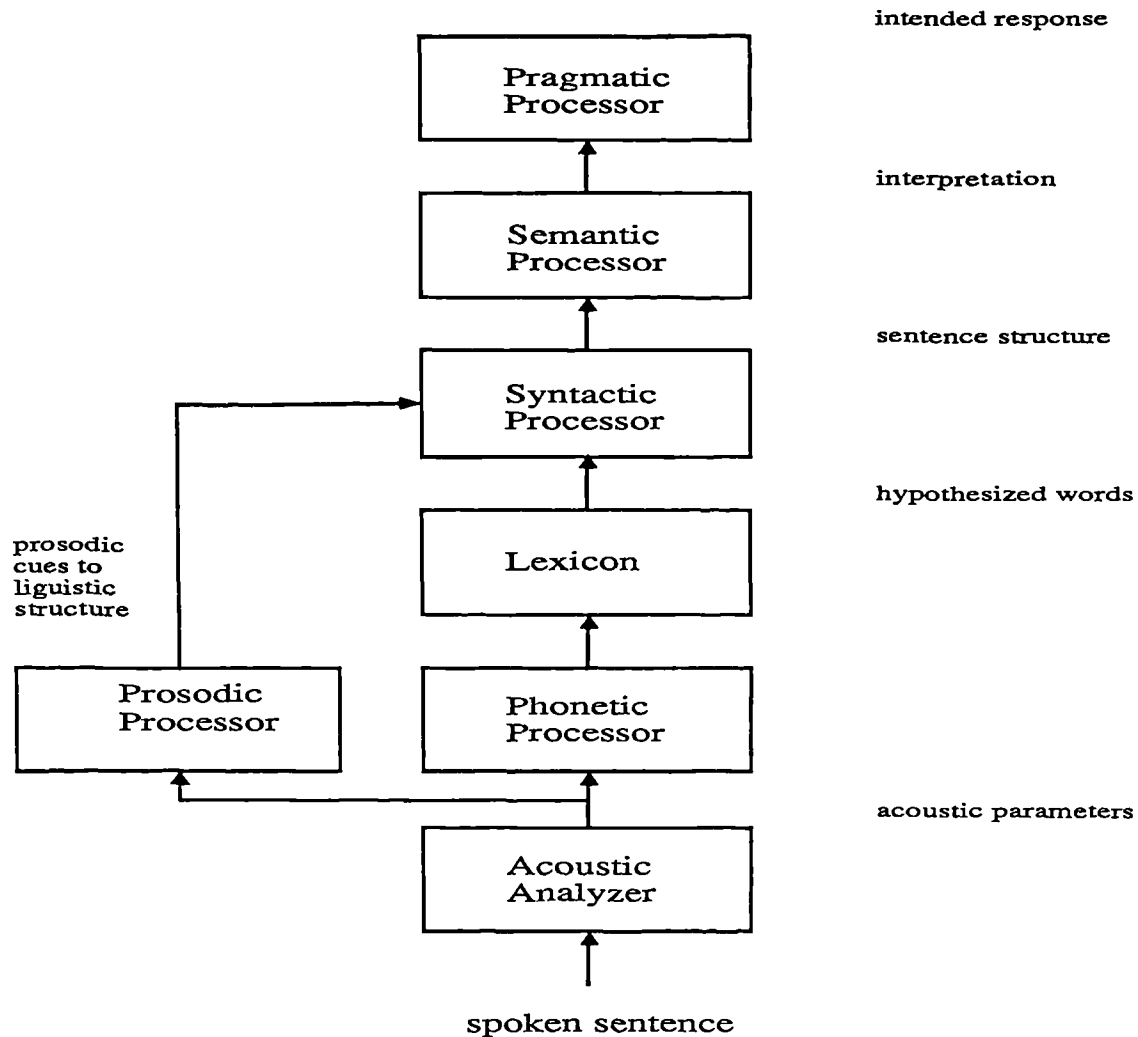


Fig. 2.10 Hierarchical speech understanding structure

The system of Figure 2.10 can be divided into two main parts, bottom-end processors and top-end processors [23]. The bottom-end (front-end) processors are the acoustic-phonetic processors, which are mainly task-independent, while the top-end processors are the other linguistic processors. The top-end processors can provide the bottom-end with constraints concerning what might be expected next.

The hierarchical structure is simple to implement, because interaction between knowledge sources are limited to those which are adjacent in the hierarchy. This limitation may be

unacceptable in practice, since more flexible interactions between knowledge sources will often be needed. For this reason, more general architectures such as the blackboard structure shown in Figure 2.11 are often employed. In this model, all data is kept in a common memory area called the blackboard, which can be accessed by any knowledge source [97]. The blackboard is a two-dimensional structure, with the dimensions being time for the start of the utterance, and level. There are eight levels representing different descriptions of the utterance according to the eight knowledge sources. However, whilst more general architectures such as these are certainly more flexible, they are also much more difficult to control. One of the advantages of this knowledge-base organisation is that each knowledge source can be modified and extended independently.

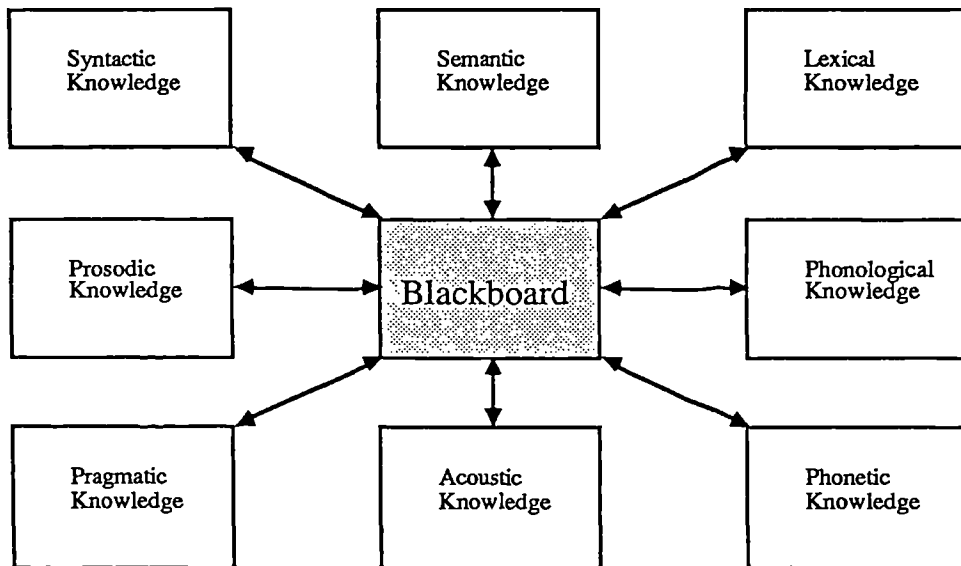


Fig. 2.11 Blackboard speech understanding structure

2.5.4 Remarks on Continuous Speech Recognition

The change in terminology from speech recognition to speech understanding in the past decade reflects a departure from the view that speech can be recognised by machine, and the acceptance of the view that linguistic knowledge sources must be used to recognise an utterance. The problem with knowledge-based understanding systems is that they are task-dependent.

It becomes increasingly evident that progress in task-independent continuous speech recognition will depend on essentially a better acoustic-phonetic analysis at the front-end of the recognition system. An almost error free bottom-up analysis is necessary for unrestricted sentence recognition.

Spectrogram reading experiments have proven that the acoustic signal is the primary information-bearer. The high scores obtained in sentence spectrogram reading (15% error in a speaker-independent phoneme identification for expert reader [93]), suggest that there exists a great deal more phonetic information in the speech signal than was previously believed, and that such information is often explicit and can be captured by rules. The performance of spectrogram reading can be improved by incorporating suprasegmental information, which can be plotted on the same spectrogram. It is worth mentioning that the spectrogram reader used a mixture of linguistic units for recognition, which are phonemes, diphones, and sub-phonemic units [93]. This does not invalidate the use of the syllable as a unit, but it seems to suggest that greater flexibility in the choice of units may be beneficial.

2.6 Word Spotting

As its name implies, word spotting is the detection of occurrences of a given key word (or words) in a stream of continuous speech. Most successful published efforts are built around DTW used in the *pattern matching model of isolated words*.

Every word to be spotted is represented by a template, and the incoming speech stream is compared with this template. This is done by sliding the input speech samples (feature vectors) past the word template in a continuous fashion and the time-warping programme tries to find paths which align the input samples with the template. Most of the time, the paths (i.e., the possible warping functions), do not make it to the end of the template as they are terminated because of excessive costs (accumulated distance measures). When a successful match is made, the system reports the presence of the key word in the input stream. It is clearly important to make the warping process independent of the endpoints, since these are known only for the template. The process must regard every sample of the incoming speech stream as a potential starting point and attempt to grow a path from it [98]. Several templates may be used for each key word in such a system to allow for intra-speaker and multiple-speaker variations (in pronunciation), and the matching is carried out over all templates in parallel [99].

A word spotter usually operates on a noisy communication channel. Five feature sets based on linear prediction, namely normalised autocorrelation coefficients, LPC coefficients, cepstral coefficients, area functions, and pseudo formants, have been used in word spotting systems working on telephone-quality speech with wide band noise added to test the system's performance on noisy speech [100]. It has been found that pseudo-formants give the best detection probability at almost every noise level. Also, a modified hidden Markov model recogniser has been used for word Spotting [101].

The best performance of a word spotter achieved an accuracy of 90 to 95% detection of a single word in noisy speech, and almost perfect performance (no misses, or no false alarms) was achieved in the absence of noise [99].

2.7 Summary

A review of speech recognition categories and techniques has been presented in this chapter. In isolated word recognition, pattern recognition approaches or stochastic modelling (based on the whole word templates or model) work very well for a small vocabulary systems. For large vocabulary systems, the acoustic-phonetic approaches are to be used. In these approaches, the recognition units are smaller than words, which are either linguistic phonetic segments or acoustic homogeneity segments (i.e., broad phonetic segments). The use of such recognition units facilitates the utilisation of linguistic knowledge to manipulate the results of the acoustic-phonetic classification process in the recognition systems. The acoustic-phonetic approaches are also used in continuous speech recognition systems.

In this research work, an acoustic-phonetic approach, which uses a hybrid scheme of broad phonetic units for consonants and detailed phonetic units for vowels, is investigated. This approach is applied to a large vocabulary isolated word Arabic speech recognition system, as explained in the following chapters. The Arabic phonetic system is introduced in Chapter 3. Chapter 4 presents the statistical results of applying different broad phonetic classification schemes to a large vocabulary lexicon. Also, it introduces the proposed recognition model. The subsequent chapters discuss the implementation of the proposed approach.

Chapter 3

Arabic Phonetic System, Phonology and Morphology

3.1 Introduction

Arabic is a Semitic language, and it is one of the oldest living languages in the world today. It is the fifth widely used language. Arabic is the mother (spoken) language throughout the Arab world (i.e., SYRIA, JORDAN, SAUDI ARABIA, EGYPT, MOROCCO, SUDAN, etc.). Arabic alphabets (and to some extent its syllabic types, morphological and syntactic structure), are used in several languages, such as Persian, Urdu and Malay.

Standard Arabic is the language of communication in official discourse, teaching, religious activities, and literature. Standard Arabic has basically 35 phonemes, of which six are vowels, and twenty-nine are consonants. Before going into details of the Arabic phonetic system, a brief description of the vocal mechanism is provided.

3.2 The Human Vocal Mechanism

Figure 3.1 illustrates a schematic cross-sectional view of the human vocal system. The vocal tract is a non-uniform tube in cross-sectional area. It extends from the glottis (i.e., the opening between the vocal cords) to the lips, and varies in shape as a function of time. In the average male, the total length of the vocal tract is about 17 cm. The cross-sectional area of the vocal tract varies along its length, from a complete closure to about 20 cm², as determined by the movement of the lips, jaw, tongue, and velum [11, 102]. The nasal cavity which begins at the velum and terminates at the nostrils can be coupled to the vocal tract by the action of the velum to produce nasal sounds. During the generation of non-nasal sounds, the velum seals off the vocal tract from the nasal cavity.

Speech sounds can be classified into three distinct classes according to their mode of excitation, namely voiced, unvoiced (voiceless), and plosive sounds. Voiced sounds are

produced by forcing air through the glottis with the tension of the vocal cords adjusted so they vibrate in a relaxation oscillation, thereby producing quasi-periodic pulses of the air which excite the vocal tract. Vowels are an example of voiced sounds. Unvoiced sounds are generated by forming a constriction at some point in the vocal tract (usually towards the mouth end), and forcing air through the constriction to produce turbulence flow. This creates a broad spectrum noise source to excite the vocal tract. Plosive sounds result from making a complete closure, building up pressure behind the closure, and abruptly releasing it.

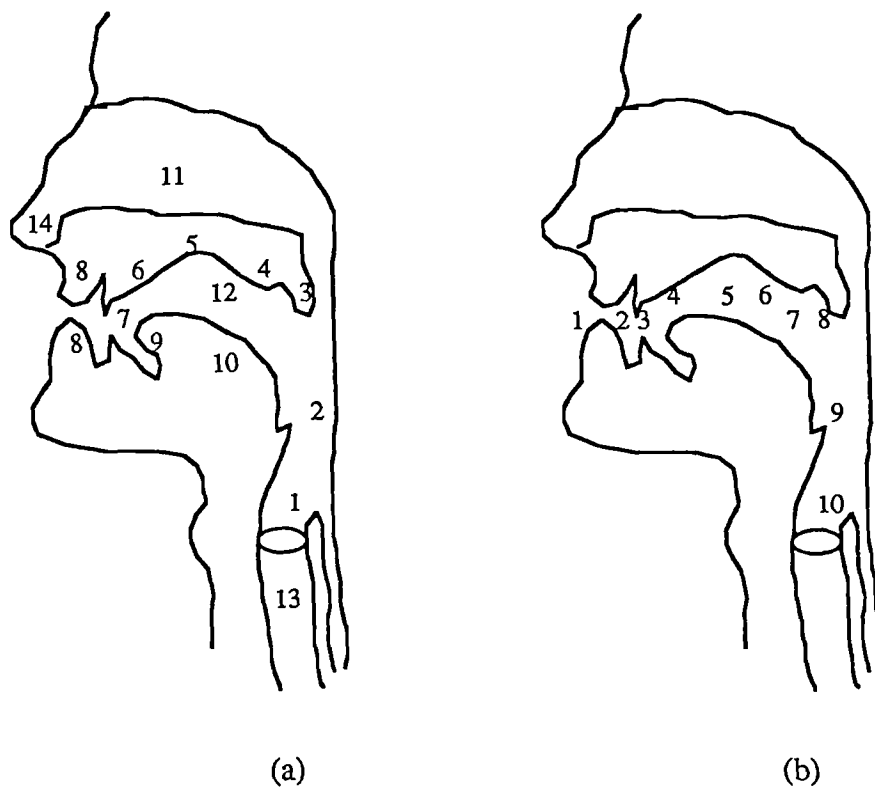


Fig. 3.1 A Schematic cross-sectional view of the human vocal system

- (a) Speech Articulators: 1) vocal cords, 2) pharynx, 3) velum, 4) soft palate, 5) hard palate 6) alveolar ridge, 7) teeth, 9) tongue tip, 10) tongue, 11) nasal cavity, 12) oral cavity, 13) trachea, 14) nostrils, 8) Lips
- (b) Places of Articulation: 1) bilabial, 2) labiodental, 3) interdental, 4) alveodental, 5) alveolar, 6) palatal, 7) velar, 8) uvular, 9) pharyngeal, 10) glottal.

The resonant frequencies of the vocal tract tube are called formant frequencies or simply formants (see Figure 2.1). Formants depend upon the shape and dimension of the vocal tract. Different sounds are formed by varying the shape of the vocal tract. Thus, the spectral properties of the speech sounds vary with time as the vocal tract shape varies. The time-varying spectral characteristics of the speech signal can be graphically displayed through the use of the sound spectrograph [13]. The spectrograph produces a two-dimensional pattern called a spectrogram, in which the vertical dimension corresponds to frequency and the horizontal dimension to time. The darkness of the pattern is proportional to signal energy, therefore we could call it a three-dimensional pattern (see Figure 3.2).

3.3 Articulatory Phonetics

The goal of articulatory phonetics is to describe speech sounds in terms of the positions of the vocal organs, and to provide a common notation for linguists.

Most languages, including Arabic, can be described in terms of a set of distinctive sounds, or phonemes. The conventional division of phonemes is into vowels and consonants.

a. Consonants

Consonants are relatively easy to define in anatomical terms. Most consonants are describable by a few recognised features, principally:

- Point of articulation
- Manner of articulation
- Voicing

i. Point of Articulation

The point of articulation is the location of the principal constriction in the vocal tract, defined usually in terms of participating organs. Table 3.1 gives a list of the principal points of articulation (relating to Arabic consonants), and the names given to the corresponding consonant.

| Name | Description |
|-------------|---|
| Bilabial | Between lips. |
| labiodental | Between lower lip and upper teeth. |
| Interdental | Tip of tongue between teeth. |
| Alveodental | Tip of tongue on gums. |
| Alveolar | Tip of tongue on alveolar ridge. |
| Palatal | Middle of tongue in contact with the hard palate. |
| Velar | Back of tongue on soft palate. |
| Uvular | Back of tongue touches or is near to the velum. |
| Pharyngeal | Root of tongue constricting oral pharynx. |
| Glottal | Between vocal cords. |

Table 3.1 Principal points of articulation

| Name | Description |
|-----------|---|
| Plosive | Vocal tract shut off at point of articulation. This is also called 'Stop'. |
| Fricative | Vocal tract partly open at points of articulation, and turbulent noise created at point of articulation. |
| Nasal | Vocal tract closed at point of articulation and velum open. |
| Semivowel | Vocal tract partly open at point of articulation without turbulence. |
| Trill | Oscillatory opening and closure at point of articulation. |
| Lateral | Vocal tract closed at point of articulation but open at sides. |

Table 3.2 Principal categories of articulation

ii. Manner of Articulation

Manner of articulation refers to the degree of constriction at the point of articulation, and the manner of release into the following sounds. Table 3.2 gives the principal manner of articulation categories of Arabic consonants and their names. Affricate consonants (such as /tʃ/) do not exist in the Arabic language, and are not included in Table 3.2. Usually the trill and lateral categories are called 'Liquid', since they are similar to vowels but are usually a few decibels weaker.

iii. Voicing

This indicates the presence or absence of vibration of vocal cords, and it is also called 'Phonation'.

b. Vowels

The mouth cavity is usually wide when pronouncing vowels, while when uttering consonants, there is often a constriction or even a closure at some point along the vocal tract. When making vowel sounds, the tongue never touches another organ, hence there is no place of articulation. Vowels are described in terms of position as follows:

- Tongue high or low (i.e., degree of constriction)
- Tongue front, back, or central.
- Lips rounded or unrounded.
- Nasalised or unnasalised.

'High or low' and 'front or back' refer roughly to the highest position of the tongue (i.e., tongue hump position). Some authors use the terms 'closed' or 'open' (to describe the mouth cavity), instead of 'high' or 'low' respectively. 'Front' is towards the lips and 'back' is towards the pharynx. In nasalised vowels, the velum is open, so that sound passes through the nasal cavity as well as the mouth, while in unnasalised vowels, the velum is shut and the sound passes through the mouth only.

3.4 Vowels

The Arabic language has six vowels, consisting of three short*¹ vowels :

- /a/ a short low central unrounded vowel
- /u/ a short high back rounded vowel
- /i/ a short high front unrounded vowel

which contrast phonetically with their long*² counterparts /aa/, /uu/, and /ii/, where phonemic length is indicated by writing the vowel twice. The duration of long vowels seem to be twice (or more) the length of short vowels, because they are usually stressed and carefully uttered in a speech sequence.

Figure 3.2 displays spectrograms of the six vowels. In this Figure, vowels are recorded as isolated utterances in order to have steady-state formants without consonant-vowel transitions. However, it is observed that almost all vowels in isolation seem to have some sort of abrupt initiation (the amount of which varies from one vowel to another).

The frequencies of the first two formants (F1 and F2), of both short and long vowels were measured as pronounced by the author. Figure 3.3 indicates the location of the vowels according to their formant measurements. Two triangles have been constructed to enclose the short and long vowels, in accordance with the classical concept of the vowel triangle [102]. The broken triangle corresponds to the short vowels and the solid triangle to the long vowels. In fact, for vowels uttered within a consonantal context, the differences between formants for short and long vowels are undistinguishable.

*1 In Arabic (and in general for all semitic languages), the three short vowels are not written as separate letters, but they are actually written as diacritic marks below or above a consonant. Usually diacritics (short vowels) are not written in normal writing, and they are identified from the contextual and grammatical structure of the sentence. The vowels /a/, /u/, and /i/ are called in Arabic 'Fatha', 'Damma', and 'Kasra' respectively.

*2 The three long vowels /aa/, /uu/, and /ii/ are written as separate letters, and called 'Alif maddiyah', 'Waw maddiyah', and 'Ya? maddiyah', respectively.

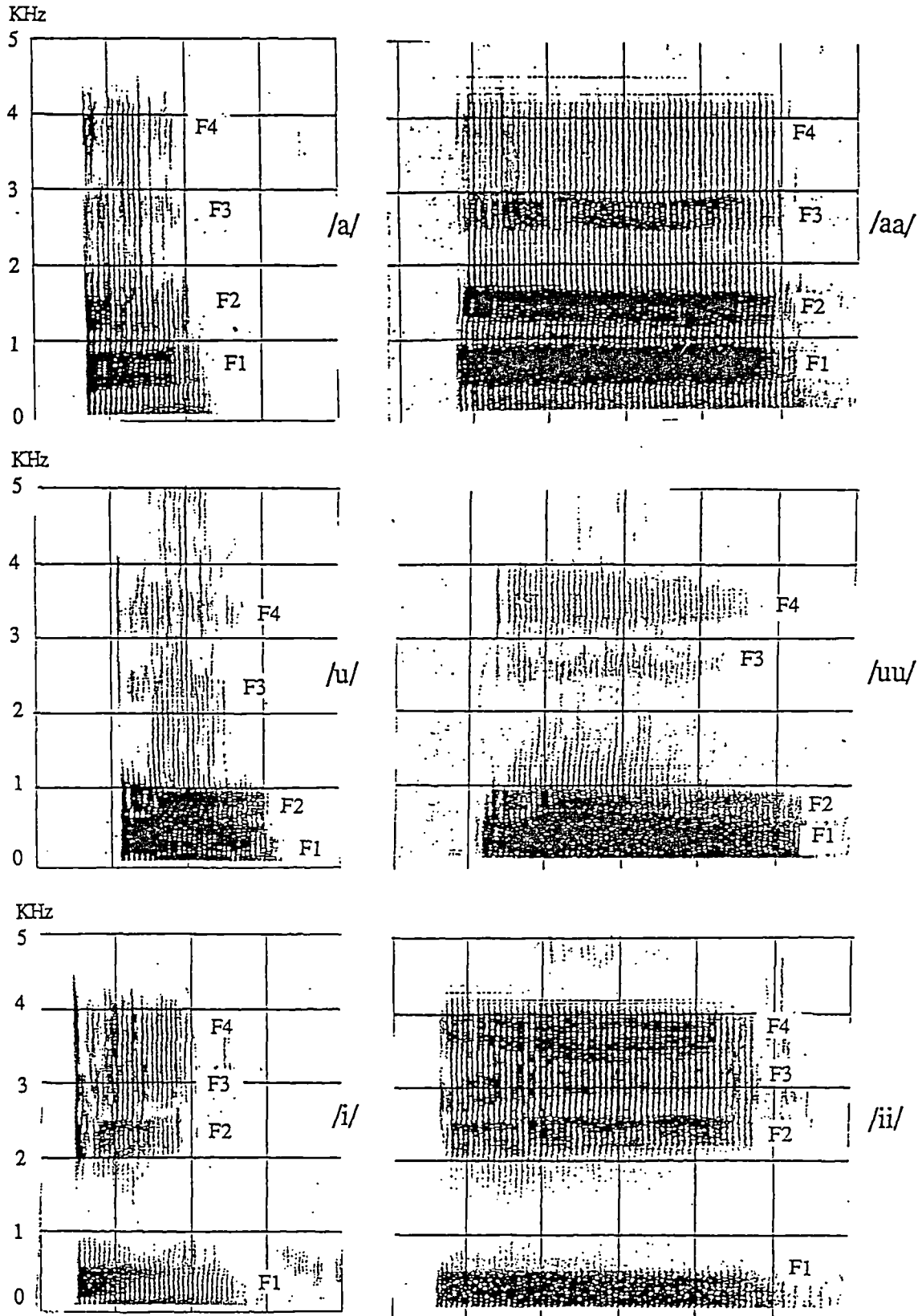


Fig. 3.2 Vowel spectrograms (short and long)

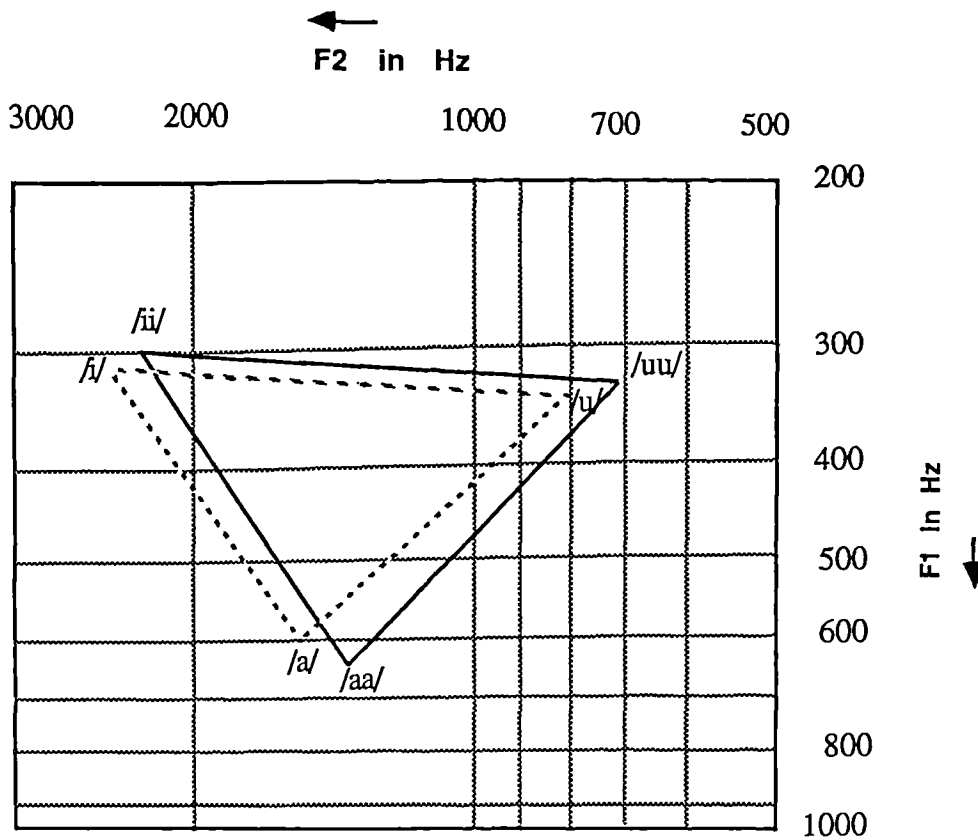


Fig. 3.3 Vowel triangle

3.5 Consonants

Arabic has basically 29 consonants. Table 3.3 shows a tentative chart of the standard Arabic phonetic system (excluding vowels) [2, 103, 104]. In this table, consonants are categorised according to their place of articulation, manner of articulation, voiced or unvoiced, and pharyngealised or non-pharyngealised. The Arabic consonantal system differs from the Latin one, primarily to the presence of pharyngealised (or emphatic), uvular, pharyngeal, and glottal phonemes. The international phonetic alphabet (IPA) has been used to describe most of the consonants in this table [105]. IPA uses the superimposed Tilde /~/, to indicate pharyngealization. To facilitate printing the symbol /~/ has been replaced by underlining the pharyngealised consonants comparable with their plain counterparts as follows:

| | |
|--------------------------|---|
| non-pharyngealised (IPA) | d , k , l , s , t , ð |
| pharyngealised (IPA) | ḏ , ḵ , ṭ , ṣ , ṫ , Ṫ |
| in Table 3.3 | <u>d</u> , <u>k</u> , <u>l</u> , <u>s</u> , <u>t</u> , <u>ð</u> |

| | | | Bilabial | Labiodental | Interdental | Alveodental | Alveolar | Palatal | Velar | Uvular | Pharyngeal | Glottal |
|-----------|----|----|----------|-------------|-------------|-------------|----------|---------|-------|--------|------------|---------|
| Plosive | vo | ph | | | | <u>d</u> | | | | | | |
| | | | b | | | d | | | | | | |
| | uv | ph | | | | <u>t</u> | | | | k | | |
| | | | | | | t | | | k | | | |
| Fricative | vo | ph | | | <u>ð</u> | | | | | | | |
| | | | | | ð | z | ʒ | | | ɣ | ʕ | |
| | uv | ph | | | | <u>s</u> | | | | | | |
| | | | | f | θ | s | ʃ | x | | | ħ | h |
| Nasal | vo | | m | | n | | | | | | | |
| Liquid | vo | ph | | | | <u>l</u> | | | | | | |
| | | | | | | l,r | | | | | | |
| Semivowel | vo | | w | | | | | j | | | | |

Table 3.3 The Arabic phonetic system (consonants)
(‘vo’: voiced, ‘uv’: unvoiced, ‘ph’: pharyngealised)

In general phonetic terms, pharyngealization or emphasis has been described as a rearward movement of the back of the tongue towards the back wall of the pharynx. The result of this movement is a vocal tract shape with an increased oral cavity (between the surface of the tongue and hard palate), and a reduced pharyngeal cavity above the epiglottis compared to non-pharyngealized counterparts. Whenever a pharyngealised consonant occurs within a syllable, the whole syllable, phonetically, is pharyngealised. Also, the pharyngealization phenomenon is not confined within the syllable boundary but

it may (or may not) have an influence on the neighbouring syllables. This puts the consonantal phonemes immediately preceding and following such a consonant in free variation (i.e., pharyngealised or non-pharyngealized). The consonant /l/ is used only in one word in the standard Arabic language, which is 'ʔallaah' (God), therefore, by avoiding this word we could confine the pharyngealised consonants to the five consonants (/d/, /t/, /s/, /k/, and /θ/).

Regarding the voiced and unvoiced description of consonants in Table 3.3, we should mention here that voiced consonants at word-final position are in free variation (i.e., voiced or unvoiced). Unvoiced consonants, intervocalic, are also in free variation [103].

3.5.1 Consonant Classes

In Table 3.3, consonants are divided into five classes according to their manner of articulation, namely plosives, fricatives, nasals, liquids, and semivowels. Each class may be divided into two sub-classes: voiced and unvoiced, and each sub-class may also be divided into two branches: pharyngealised and non-pharyngealized.

The speech sound appears on the spectrogram as follows:

- Voiced plosives appear as a voice bar along the baseline followed by a sudden burst noise.
- Unvoiced plosives appear as a gap followed by a sudden burst noise.
- Unvoiced fricatives usually possess a high frequency random noise.
- Voiced fricatives usually possess weak resonance structures appearing as shadows of weak formants with little noise intervening. The strongest of these formant structures, indicating the voicing, appears along the baseline.
- Nasals appear as voice bars along the baseline (which differentiate the nasals from the vowels and other vowel-like sounds), and possess weak resonances that appear as formant structures.
- Liquids are sonorant consonants and have spectra very similar to vowels. The /r/, sometimes, shows distinct formant structures, interrupted by a vertical sharp gap with short duration (Trill).
- Semivowels possess acoustical characteristics more similar to those of the vowels than any other consonantal groups. They possess vowel-like formant structures.

Spectrograms of all possible combinations for consonant-vowel pairs (29 consonants and three long vowels) are shown in Figures 3.4 to 3.6 and in appendix B.

3.5.2 Vowel in Pharyngealized Context

This section demonstrates some of the acoustical properties of vowels when they are next to pharyngealised consonants. Figures 3.4 to 3.6 display spectrograms for all combinations of the five pharyngealised consonants and their non-pharyngealised counterparts with the three vowel types.

Figure 3.4 displays spectrograms for the vowel /i/ next to the consonants /d/, /t/, /s/, /k/, and /θ/, and their pharyngealised counterparts. When pronouncing the front vowel /i/, the hump of the tongue is toward the lips, but when uttering a pharyngealised consonant, the whole body of the tongue is in a backward movement and the back of the tongue is close to the back wall of the pharynx. When /i/ (or /i/) comes next to pharyngealised consonants, this means that the tongue has to move from its back position to its front position. This phenomenon is translated acoustically by longer transitional time for the second formant (F2) in the vowel part, as is illustrated in all cases of Figure 3.4. This transition takes about one-third of vowel's duration for the vowel /i/ and about one-fifth for the vowel /ii/. Also, we notice that the onsets of F1 and F2 for the vowel /ii/ are influenced by the pharyngealised consonants, where the onset of F2 is lowered and the onset of F1 is raised compared to those in the non-pharyngealised context.

Figure 3.5 displays spectrograms for the vowel /aa/ next to the consonant /d/, /t/, /s/, /k/, and /θ/, and their pharyngealised counterparts. From these spectrograms, we notice that not only the onsets of formants are affected, but also the formant steady states. In general F1 and F2 for the vowel /aa/ (or /a/) move closer to each other in the pharyngealised context compared to the non-pharyngealized context. We also notice, that F3 and F4 move slightly closer to each other (see Figure 3.5).

Figure 3.6 shows spectrograms for the vowel /uu/ next to the consonants /d/, /t/, /s/, /k/, and /θ/, and their pharyngealised counterparts. These spectrograms show that the vowel's formant onsets as well as its formant frequencies are influenced in a pharyngealised context. This influence is not as clear as in the case of the vowel /aa/.

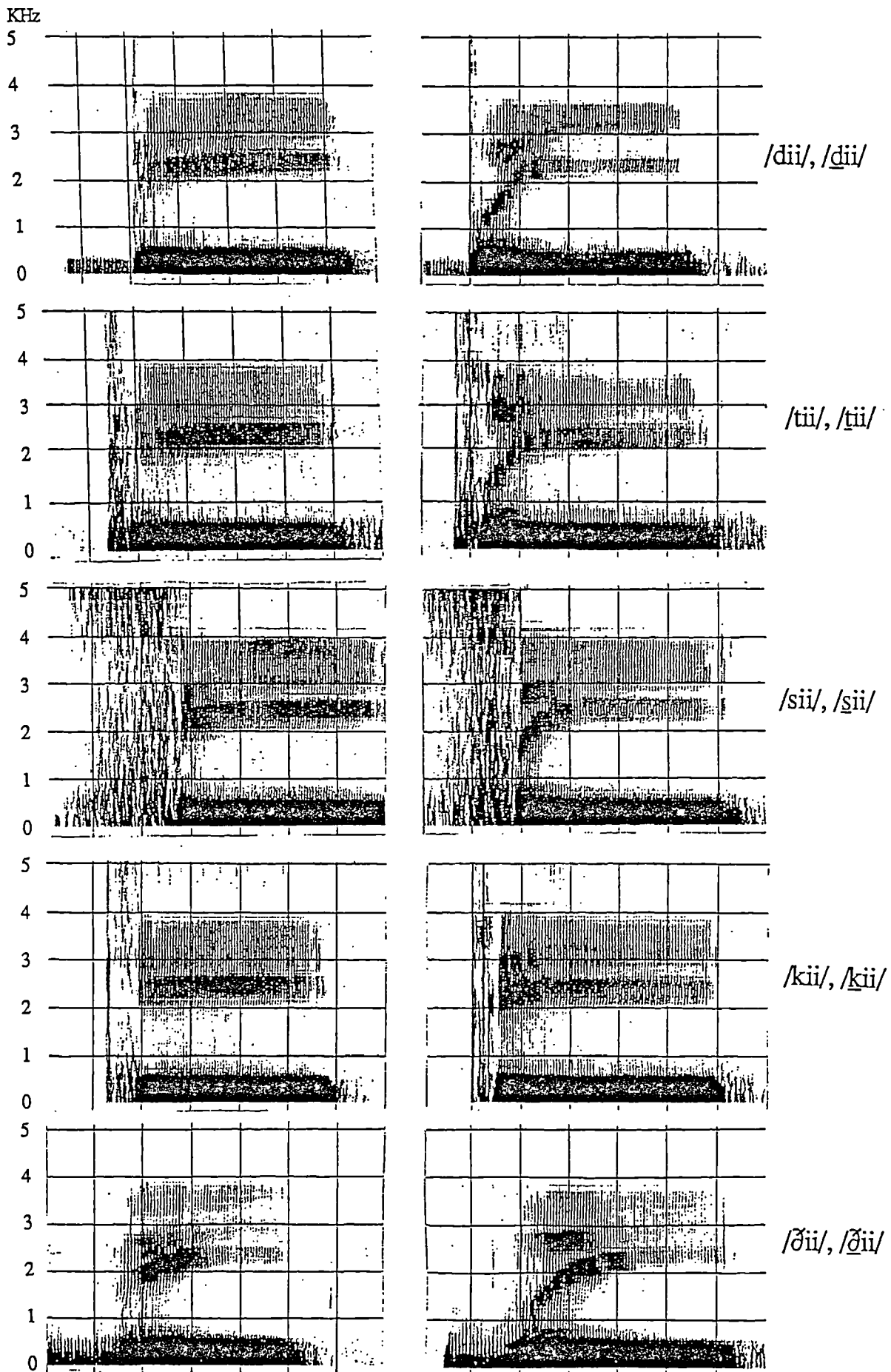


Fig. 3.4 Consonant-vowel spectrograms (vowel /ii/)

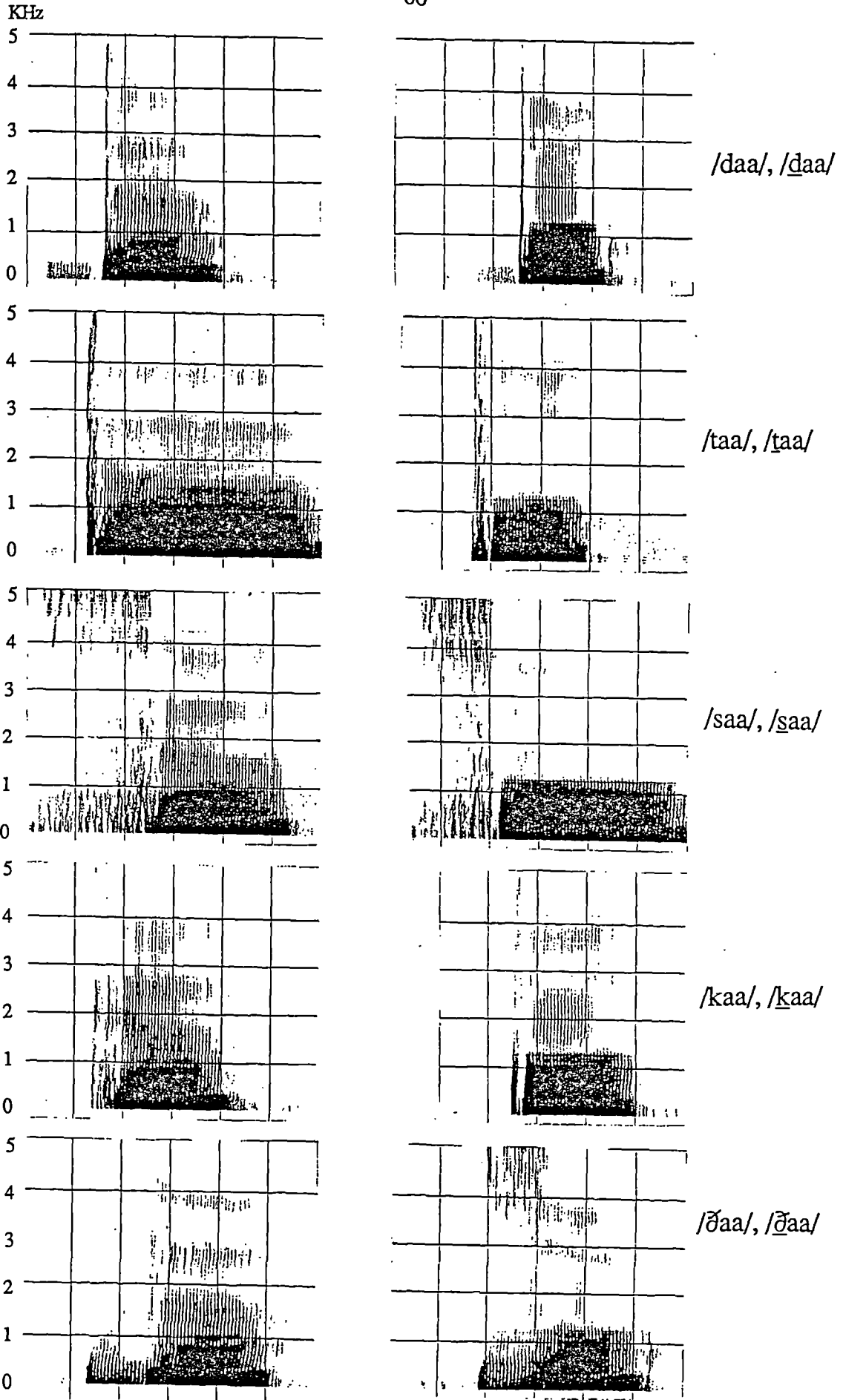


Fig. 3.5 Consonant-vowel spectrograms (vowel /aa/)

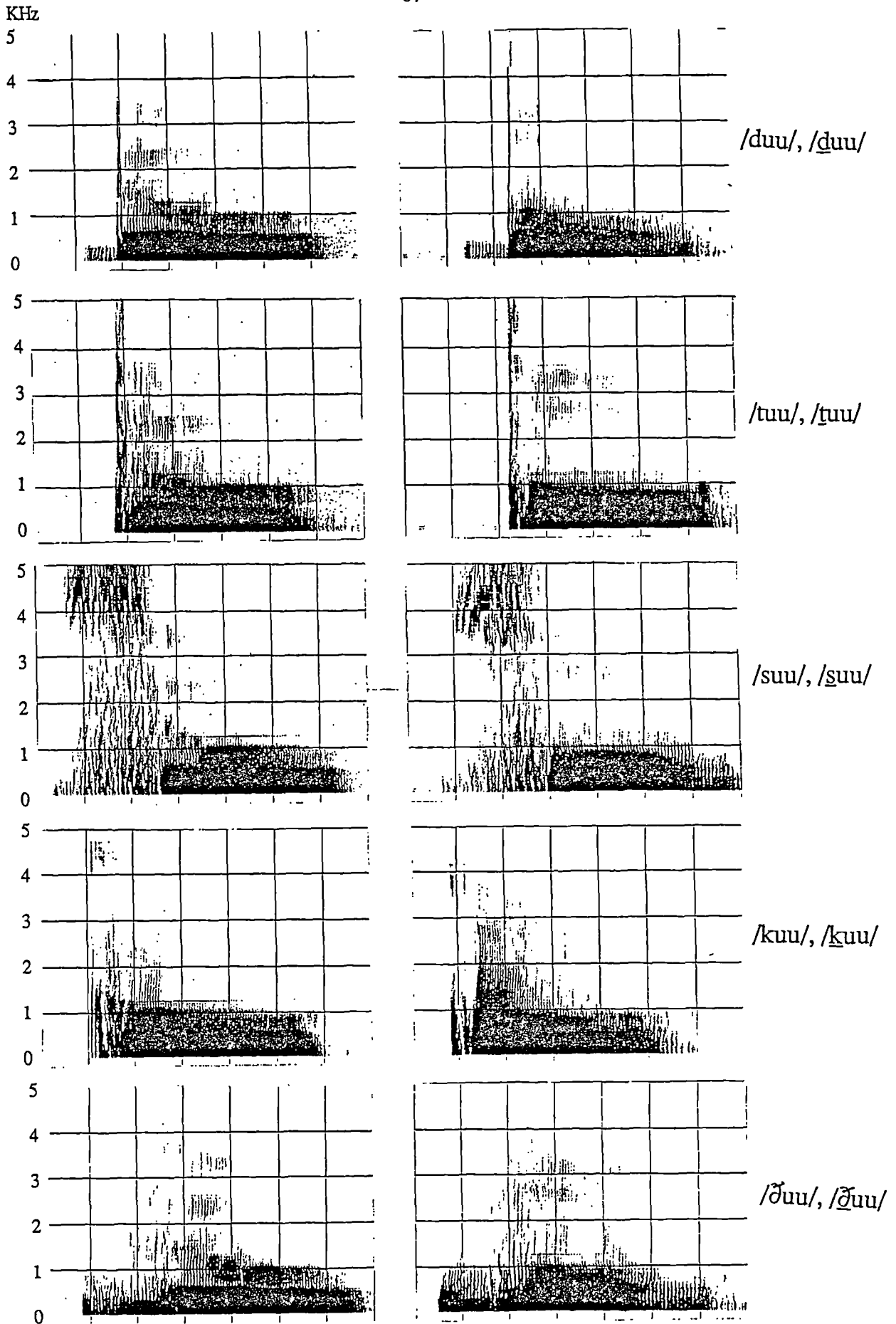


Fig. 3.6 Consonant-vowel spectrograms (vowel /uu/)

3.6 Distribution of Vowel and Consonant Classes

Statistical studies have been carried out to investigate the distribution of vowel and consonant classes using the lexical database mentioned in Chapter 1. This database contains two lexicon, the first one contains the most common 3,000 words in the Arabic language and the second contains 10,000 randomly chosen words. Words are described in phonemic forms.

3.6.1 Distribution of Vowels

Table 3.4 displays the frequencies of occurrence of the six vowels for the first lexicon, and Table 3.5 shows the results for the second lexicon [106].

| | no. of occurrence | % of phoneme | % of vowels |
|-----------------------|-------------------|--------------|-------------|
| a | 3931 | 21.65 | 53.48 |
| u | 751 | 4.14 | 10.22 |
| i | 970 | 5.34 | 13.20 |
| aa | 1227 | 6.76 | 16.69 |
| uu | 169 | 0.93 | 2.30 |
| i i | 302 | 1.66 | 4.11 |
| vowels | 7350 | 40.48 | 100. |
| consonants | 10809 | 59.52 | |
| total no. of phonemes | 18159 | | |

Table 3.4 Distribution of vowels for the 3,000-word lexicon

| | no. of occurrence | % of phoneme | % of vowels |
|-----------------------|-------------------|--------------|-------------|
| a | 15417 | 20.32 | 47.33 |
| u | 4744 | 6.25 | 14.57 |
| i | 5791 | 7.63 | 17.78 |
| aa | 4417 | 5.82 | 13.56 |
| uu | 1058 | 1.39 | 3.25 |
| i i | 1143 | 1.51 | 3.51 |
| vowels | 32570 | 42.92 | 100. |
| consonants | 43305 | 57.08 | |
| total no. of phonemes | 75875 | | |

Table 3.5 Distribution of vowels for the 10,000-word lexicon

Table 3.5 shows that the six vowels represent about 43% of the total number of phonemes, while the 29 consonants represent about 57%. The short vowel /a/ has the highest percentage (47%) among the six vowels, and the vowels /a/ and /aa/ together

represent about 60% of the total number of vowels. The vowels /u/, /i/ and /aa/ have almost similar frequencies of occurrence.

Table 3.4 gives slightly different results, where the amount of data used here is less than that of Table 3.5. The importance of Table 3.4 is, that it gives results for the most common 3,000 words in the language, which were extracted from about one million words, and they represent about 84% of all words [10]. Unfortunately, we did not load the frequencies of occurrence of these 3,000 words into our computer, otherwise we would have had a wider range of statistical results covering about one million words. Nevertheless, the results are quite impressive, especially the frequency of occurrences for the vowels /a/ and /aa/.

3.6.2 Distribution of Consonant Classes

The consonant classes, in terms of manner of articulation as introduced in Section 3.5, have been considered in this statistical study.

Tables 3.6 and 3.7 report the distribution of consonant classes over the two lexicons in the lexical database. By comparing the two tables, we notice that there is no substantial difference between the statistical results of the two sets of words, apart from the decrease in the percentage of the fricative consonants by about 5% (from 32.5% to 27.83% with respect to the total number of consonants), and the increase of nasals by about 4.5% (from 15.52% to 19%). Regarding voicing, about 76% of the total number of phonemes are voiced and 24% are unvoiced (for the 10,000 word lexicon), whereas about 58% of the consonants are voiced consonants.

It is believed, that this study is the first of its type concerning Arabic speech processing, since all previous statistical studies [107, 108] have concentrated on the written aspects of the Arabic language. Two studies have dealt with the frequencies of occurrence for vowels and consonants [109, 110]. The difference between these studies and the current work lie in the databases. Words occur only once in the lexicon in our data base while the other studies used normal text, i.e., words may occur more than once in the data base. Nevertheless, the statistical results for vowels are somewhat similar in the three studies.

| | no. of occurrence | % of phoneme | % of class | | no. of occurrence | % of phoneme | % of class | | no. of occurrence | % of phoneme | % of class |
|------------|-------------------|--------------|------------|----|-------------------|--------------|------------|----|-------------------|--------------|------------|
| Plosives | 3461 | 19.06 | 32.02 | vo | 1185 | 6.53 | 10.96 | ph | 138 | 0.76 | 1.28 |
| | | | | | | | | | 1047 | 5.77 | 9.68 |
| | | | | uv | 2276 | 12.53 | 21.06 | ph | 643 | 3.54 | 5.95 |
| | | | | | | | | | 1633 | 8.99 | 15.11 |
| Fricatives | 3513 | 19.35 | 32.50 | vo | 1150 | 6.34 | 10.64 | ph | 70 | 0.39 | 0.65 |
| | | | | | | | | | 1080 | 5.95 | 9.99 |
| | | | | uv | 2363 | 13.01 | 21.86 | ph | 211 | 1.16 | 1.95 |
| | | | | | | | | | 2152 | 11.85 | 19.91 |
| Nasals | 1678 | 9.24 | 15.52 | | | | | | | | |
| Liquids | 1588 | 8.74 | 14.69 | | | | | | | | |
| Semivowels | 569 | 3.13 | 5.27 | | | | | | | | |
| | 10809 | 59.52 | 100. | | | | | | | | |

Table 3.6 Distribution of consonant classes for the 3,000-word lexicon ('vo': voiced and 'uv': unvoiced)

| | no. of occurrence | % of phoneme | % of class | | no. of occurrence | % of phoneme | % of class | | no. of occurrence | % of phoneme | % of class |
|------------|-------------------|--------------|------------|----|-------------------|--------------|------------|----|-------------------|--------------|------------|
| Plosives | 14436 | 19.03 | 33.33 | vo | 4139 | 5.46 | 9.56 | ph | 500 | 0.66 | 1.16 |
| | | | | | | | | | 3639 | 4.80 | 8.40 |
| | | | | uv | 10297 | 13.57 | 23.72 | ph | 2398 | 3.16 | 5.53 |
| | | | | | | | | | 7899 | 10.41 | 18.24 |
| Fricatives | 12051 | 15.88 | 27.83 | vo | 4094 | 5.39 | 9.45 | ph | 268 | 0.35 | 0.62 |
| | | | | | | | | | 3826 | 5.04 | 8.83 |
| | | | | uv | 7957 | 10.49 | 18.38 | ph | 698 | 0.92 | 1.62 |
| | | | | | | | | | 7259 | 9.57 | 16.76 |
| Nasals | 8226 | 10.84 | 19.00 | | | | | | | | |
| Liquids | 5637 | 7.43 | 13.02 | | | | | | | | |
| Semivowels | 2955 | 3.89 | 6.82 | | | | | | | | |
| | 43305 | 57.08 | 100. | | | | | | | | |

Table 3.7 Distribution of consonant classes for the 10,000-word lexicon ('vo': voiced and 'uv': unvoiced)

3.7 Syllabic Types and Structure (phonology)

Every language has its own syllabic system that characterises it from other languages. For many European languages such as English, French, and German, the nucleus of a syllable is necessarily a vowel. Arabic has also the same property. In English, the vowel may be preceded by a cluster of consonants (which is called the leading consonant cluster), and followed by a cluster of consonants (which is called the trailing consonant cluster). For example, the words 'stop' and 'street' are monosyllabic words; in the first word the leading or initial consonant cluster has two consonants, while in the second word it has three consonants. In some cases, a leading or trailing cluster may contain a single consonant or even no consonant at all. The latter case applies to syllables starting or ending with a vowel or diphthong, e.g., 'in' and 'to'. In the Arabic language, the vowel must be preceded by only one consonant and it may be followed by two consonants or one consonant, or no consonant at all.

3.7.1 Syllabic Types

Arabic language uses three main syllabic types (or patterns), namely: CV, CVC, and CVCC, where /C/ refers to a consonant and /V/ refers to a vowel. Some linguists distinguish between syllable types on the basis of having a long or a short vowel [111], therefore they define five syllable types:

- | | | |
|--------|---------|---------|
| 1) CV | 3) CVC | 5) CVCC |
| 2) CVV | 4) CVVC | |

where /VV/ refers to a long vowel. Note that the last type which has two trailing consonants does not have a long vowel. The first three types may occur at word-initial, word-medial, or word-final position, and they are more frequent in Arabic words. The fourth type occurs mainly at the word-final position, but it may occur at the word-initial or word-medial position only when the trailing consonant of this syllable /CVVC/ is geminated (i.e., the same consonant appears as a trailing consonant of the current syllable and as a leading consonant for the following syllable in a word). For example, in the word:

'maaddah' /CVVC-CVC/ (substance)

which has two syllables, the consonant /d/ is geminated, i.e., the first /d/ belongs to the

first syllable, while the second /d/ belongs to the second syllable. The fifth syllabic type occurs only at the word-final position or in isolation (monosyllabic words). Examples of monosyllabic words are as follows:

| | | |
|--------|--------|---------|
| 'nahr' | /CVCC/ | (river) |
| 'jawm' | /CVCC/ | (day) |

Two-consonant clusters may appear at the word-medial position, and only before a pause at the word-final position.

Arabic words are constructed mostly from three and/or four syllables, and infrequently words are made of five syllables or more. However, Arabic is characterised by a well-defined syllabic structure.

3.7.2 Phonological Constraints

Some phonological constraints regarding the syllable types and their occurrences are summarised as follows:

- Each syllable must have a leading consonant in Arabic, so words always start with a consonant.
- Whenever a long vowel occurs in the word-initial or word-medial position, the related syllable (containing this vowel) must be an open syllable (of the type /CVV/), unless the trailing consonant is geminated.
- The maximum allowable number of consonants in a consonant cluster is only two. This may appear in the word-medial position or in the word-final position. In the former case, the first consonant is considered as the trailing consonant of the current syllable and the second consonant is considered as the leading consonant of the following syllable. In the latter case, both consonants belong to the current syllable (type: /CVCC/).

These simple phonological constraints are very important. They have been exploited in the segmentation procedure in the recognition system as will be shown in Chapter 7.

Another important phonological constraint is the allowable consonant structure, or in other words, the consonant association and dissociation. There are cacophonous

constraints imposed on consonant association, but some consonantal combinations are not used in the language [112]. The cacophonous constraints are related to those consonants having identical or adjacent place of articulation. For instance:

- The pharyngeal consonant /ħ/ can not associate (in the same consonant cluster and maybe in the same word) with the pharyngeal consonant /ʕ/, or with the glottal consonant /h/.
- The uvular /k/ can not associate with the velar /g/.
- The interdental /θ/ can not associate with the alveolar /s/, /z/, and /d/.

Table 3.8 shows those consonants which may not be combined in one consonant cluster or may not even occur in the same syllable or in the word [113].

3.7.3 Gemination

Gemination involves a longer closure of the plosive consonant and prolongation of other consonants. All consonants in Arabic have both short and long (geminated) phonetic realisations. For example, the following words have geminated consonants:

| | | |
|--------------------|--------------|----------------------|
| 'mattana' | /CVC-CV-CV/ | (cause to be strong) |
| ' <u>k</u> attala' | /CVC-CV-CV/ | (he slaughtered) |
| 'maaddah' | /CVVC-CVC/ | (Substance) |
| 'yassaala' | /CVC-CVV-CV/ | (washing machine) |

As far as the syllable boundaries are concerned, the first consonant is the trailing consonant of the preceding syllable and the second consonant is the leading consonant of the following syllable.

For geminated plosive consonants, the closure is longer than that for the case of normal plosive consonants. For other geminated consonants, the first consonant is not released until the second consonant is uttered, resulting in a longer duration.

In writing, a geminated consonant is written as a single consonant, and a special diacritic mark (called 'jadda') is superimposed on the consonant. However in printing, the 'jadda' is not printed, as well as other diacritic marks, and the reader has to extract the actual pronunciation from the syntactic structure of the sentence.

| phoneme | phonemes which can not associate with (follow) the phoneme in the first column |
|---------|---|
| ʔ | ʔ, ε |
| b | f |
| t | θ, θ̃, s, d, t |
| θ | θ, z, s, s̃, d, θ̃, ʃ |
| ʒ | t, t̃, k, γ |
| ħ | ε, γ, h, x |
| x | ʔ, γ, ħ, h, k |
| d | t, t̃, d, θ̃ |
| θ̃ | t, θ, z, s, ʃ, s̃, d, θ̃, d, t̃ |
| r | θ̃ |
| z | θ, s, ʃ, s̃, d, θ̃, θ̃ |
| s | θ, z, ʃ, s̃, d, θ̃ |
| ʃ | d |
| s̃ | θ, s, ʃ, d, θ̃, θ̃, z |
| d̃ | θ, s, ʃ, s̃, θ̃, θ̃, t, k |
| t̃ | t, s̃, d, θ̃, θ̃ |
| θ̃̃ | t, θ, ʒ, ħ, x, d, θ̃, z, s, ʃ, s̃, d, t̃, γ, k, k, h |
| ε | ʔ, ħ, x, γ |
| γ | ʔ, ħ, ʒ, x, ε, k |
| f | b |
| k̃ | ʒ, k |
| k | t, k |
| m | b, f |
| h | ħ, x, θ̃ |

Table 3.8 Illegal consonant clusters (combination) [113]

3.7.4 Distribution of Syllabic Patterns

The distribution of the five syllabic types has been investigated using the lexical database. Each syllabic type can take three different vowels, yielding three different syllabic patterns. So we have 15 different syllabic patterns for the five syllabic types.

Tables 3.9 and 3.10 display the frequencies of occurrence of the fifteen syllabic patterns using the two lexicons in our database [106].

| | no. of occurrence | % of each type | pattern | % of each pattern |
|------|----------------------|-------------------|---------|-------------------------|
| ca | 2426 | 33.01 | cv | 42.38 |
| cu | 301 | 4.10 | | |
| ci | 388 | 5.28 | | |
| cac | 1421 | 19.33 | cvc | 32.85 |
| cuc | 434 | 5.91 | | |
| cic | 559 | 7.61 | | |
| cacc | 84 | 1.14 | cvcc | 1.67 |
| cucc | 16 | 0.21 | | |
| cicc | 23 | 0.31 | | |
| caa | 765 | 10.14 | cvv | 12.23 |
| cuu | 36 | 0.49 | | |
| cii | 98 | 1.33 | | |
| caac | 462 | 6.29 | cvvc | 10.87 |
| cuuc | 133 | 1.81 | | |
| ciic | 204 | 2.77 | | |

Table 3.9 Distribution of syllabic patterns for the 3,000-word lexicon

The results for both sets of words are somewhat similar, especially for the two syllabic types /CVCC/ and /CVVC/. The differences come from the fact that the syllable /CVVC/ occurs mainly in isolation (monosyllabic words), and the total number of these monosyllabic words is 123 in both lexicons (3,000 and 10,000 words). Also, the

syllable /CVVC/ occurs 799 times in the 3,000-word lexicon and 963 times in the 10,000-word lexicon. The increase in the number of words from 3,000 in the first lexicon to 10,000 in the second lexicon did not coincide with increases in the number of these syllabic types (/CVCC/ and /CVVC/). The monosyllabic words are limited in the language, and the syllable /CVCC/ is less frequent in the Arabic language, because it occurs mainly at the word-final position.

Table 3.10 gives better statistical estimation of the frequencies of occurrence of 15 syllabic patterns (where the total number of syllables under investigation is 32570). From this table, we notice that the syllables /CV/ and /CVC/ are much more frequent than others, and then comes the type /CVV/. The Type /CV/ represents about 50% of the total number of syllables, while the type /CVC/ represents about 29%. The two open syllables /CV/ and /CVV/ represent 67% of the total number of syllables. The two patterns /Ca/ and /Caa/ represent about 40% of the total number of syllables and about 60% of the total number of the syllabic types /CV/ and /CVV/.

| | no. of occurrence | % of each type | pattern | % of each pattern |
|------|----------------------|-------------------|---------|-------------------------|
| ca | 9187 | 28.20 | cv | 50.05 |
| cu | 3303 | 10.14 | | |
| ci | 3813 | 11.71 | | |
| cac | 6146 | 18.87 | cvc | 29.25 |
| cuc | 1425 | 4.38 | | |
| cic | 1955 | 6.00 | | |
| cacc | 84 | 0.26 | cvcc | 0.38 |
| cucc | 16 | 0.05 | | |
| cicc | 23 | 0.07 | | |
| caa | 3859 | 11.85 | cvv | 17.36 |
| cuu | 903 | 2.77 | | |
| cii | 893 | 2.74 | | |
| caac | 558 | 1.71 | cvvc | 2.96 |
| cuuc | 155 | 0.48 | | |
| ciic | 250 | 0.77 | | |

Table 3.10 Distribution of syllabic patterns for the 10,000-word lexicon

3.8 A Brief Description of the Arabic Morphological System

The Arabic language, like all other semitic languages, has a very systematic morphological structure, compared with Latin languages. There exist strict morphological rules which control the vocabulary structure.

a. Roots

Arabic words are morphologically derived from a shorter set of generative roots. Arabic has 11347 roots [113]. They are basically triradical, quadriradical and five-radical roots, e.g., the triradical root is made of three consonants (letters). As we mentioned earlier in this chapter, short vowels are written as diacritic marks, and are not counted in Arabic roots, therefore, triradical roots may appear in four valid syllabic structures, namely /CV-CV-CV/, /CV-CVC/, /CVC-CV/, /CVCC/. Long vowels are not used with roots, but they are used in their derivatives. The triradical roots are used much more frequently than quadriradical or five-radical ones, and they represent about 63% of all roots. The following are examples of words constructed from a triradical root:

| | | |
|----------|------------|------------|
| 'kataba' | /CV-CV-CV/ | (he wrote) |
| 'faʿala' | /CV-CV-CV/ | (he did) |

the first word has three consonants: /k/, /t/, and /b/, and three vowels (diacritic marks) /a/, and the second word has three consonants /f/, /ʿ/, and /l/. Each root can be used to generate hundreds of words according to some specific patterns.

b. Morphological Patterns

Arabic words are classified into three main categories, namely verb, noun, and tool. Examples of the last category are pronouns, preposition, and affixes. Each one of the two other categories (i.e., verb and noun) has its own sub-categories. Verb, for example, consists of two main categories: three-letter verbs and four-letter verbs, and each category contains two sub-categories, namely active verbs and passive verbs. Moreover, each sub-category also has three branches according to the tense of the verb (i.e., past, present, and imperative). Finally, each tense may have two sub-branches, namely 'mujarrad', and 'maziid', where 'maziid' means that some phonemes (one, two, or three phonemes from a pre-specified set of 10 phonemes, namely /s/, /ʔ/, /l/, /t/, /m/, /uu/, /n/,

/j/, /h/, and /aa/), are added to the original verb to extract a new morphological pattern, while 'mujarrad' means that no extra letters are added to a verb. At each final branch of the verbal categories, a list of morphological patterns may exist. Figure 3.7 shows a simplified diagram for verbal categories. Examples of three-letter active past verbs are:

| | | |
|-----------|-------------|--------------|
| 'haraba' | /CV-CV-CV/ | 'he escaped' |
| 'kasara' | /CV-CV-CV/ | 'he broke' |
| 'saraka' | /CV-CV-CV/ | 'he stole' |
| 'sarakat' | /CV-CV-CVC/ | 'she stole' |
| 'sarakuu' | /CV-CV-CVV/ | 'they stole' |

The two last example are 'maziid' verbs, where the consonant /t/ is added at the end of the word 'saraka', to indicate feminine, and the vowel /a/ is replaced by the vowel /uu/ to indicate plural. The syllabic pattern /CaCaCa/ is called a morphological pattern or balance (morphological balance, or simply balance, is the classical term used by Arab grammarians to refer to the morphological pattern). This pattern defines the syllabic structure and the actual vowels used in a word. In the balance /CaCaCa/, the consonant /C/ could be any of the 29 Arabic consonants, but the sequence of consonants is subject to phonological rules or constraints.

Apart from the morphological balances for nouns, there is a list of balances for names derived from verbs such as: present and past participle, place names, adverbs of time, machine names, comparison names (of adjectives), etc., for example:

| | | |
|------------|--------------|-------------------|
| 'laeiba' | /Ca-Ci-Ca/ | 'he played' |
| 'maleab' | /CaC-CaC/ | 'playing field' |
| 'yasala' | /Ca-Ca-Ca/ | 'he wash' |
| 'maysal' | /CaC-CaC/ | 'wash room' |
| 'yassaala' | /CaC-Caa-Ca/ | 'washing machine' |

Nouns also have similar categorical descriptions. For example, three-letter nouns are divided into two categories, namely male and female, and each of these is divided into three categories, namely singular, dual, and plural nouns. Each final branch has its own list of morphological patterns.

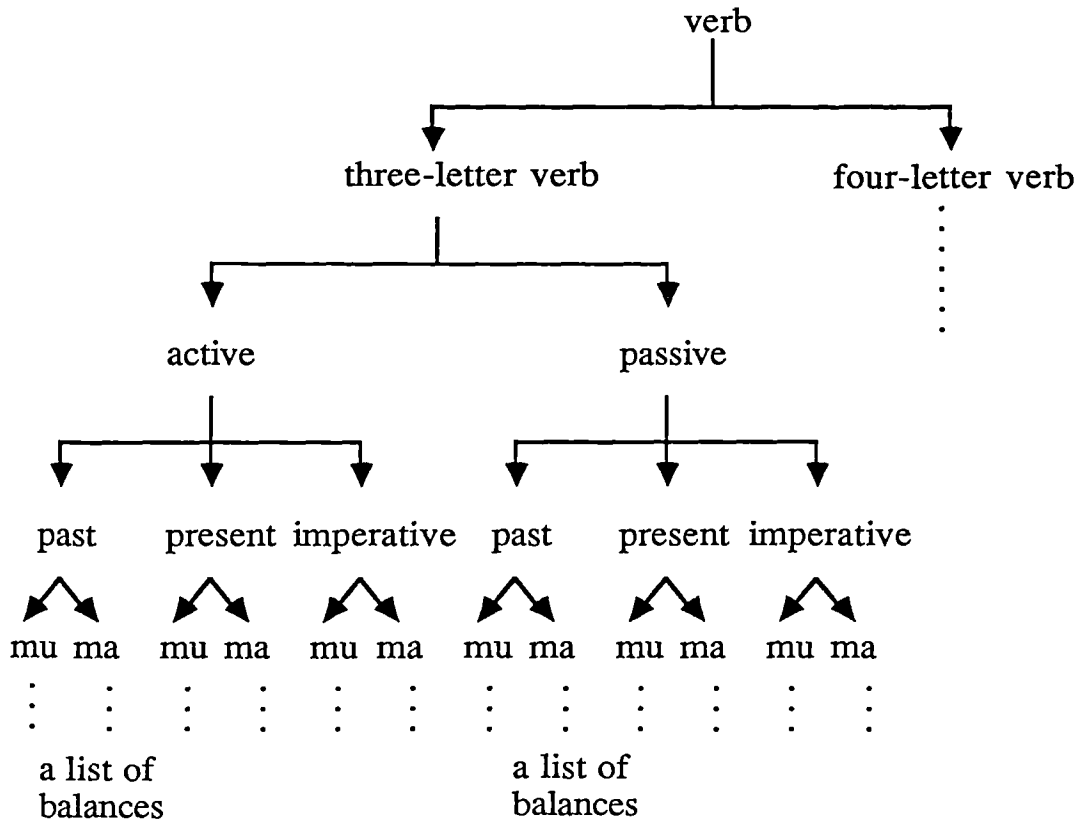


Fig. 3.7 Verb categories
 ('mu': 'mujarrad', and 'ma': 'maziid')

As a result, each word in the language (except for some non-inflectional words like pronouns, prepositions, etc.), has a morphological pattern or balance. Thus, there exist balances for all the derivatives of each root. By having a certain root and the set of balances, the derivatives of that root can easily be extracted according to certain synthesis rules. Also, the original root of a certain word can be extracted from that word according to certain analysis rules. Both analysis and synthesis rules are based on the morphological pattern.

The morphological balance can be considered as a vital tool in speech recognition, where detecting the balance is the basic step towards grammatical, syntactic and semantic analysis for the Arabic language.

3.9 Prosodic Features

Prosodic features or suprasegmental features provide information beyond phoneme boundaries and span groups of syllables or words. Prosodic information usually encompasses stress, duration, and intonation.

3.9.1 Stress

Stress reflects the degree of emphasis with which a word or a syllable is spoken. Stressed sounds are usually louder, but they are also longer and more tense. Stress tends to move vowels out toward the vertices of the vowel triangle (see Figure 3.3). Unstressed vowels tend to be pronounced closer to the neutral position (centralised formants, or vowel reduction phenomenon). Stress tends to raise pitch variation. It is quantised to three levels, namely primary stress /´/, secondary stress /˘/, and weak (unmarked).

Stress is applied to syllables, words and to longer units of speech. Individual words normally have one syllable which is stressed. Phonetically every word has an inherently-stressable syllable. This syllable receives the primary stress. Its location and distribution are affected by the number and types of syllables contained in a word. A monosyllabic word, in isolation, receives the primary stress. Disyllabic and polysyllabic words receive secondary and weak stresses in addition to the primary stress [103].

Some rules which govern lexical item stress in Arabic are as follows:

- When a word is made up of a string of /CV/ type syllables, the first syllable receives the primary stress, and the remaining syllables receive weak stress, such as:

| | | |
|----------|--------------------------|--------------|
| 'kátaba' | /C [´] V-CV-CV/ | 'he wrote' |
| 'dárasa' | /C [´] V-CV-CV/ | 'he studied' |

- When a word contains only one long syllable, this syllable receives the primary stress and the rest go unmarked, receiving weak stresses such as:

| | | |
|---------------|---|---------------|
| 'káatib' | /C [´] VV-CVC/ | 'writer' |
| 'muεállimuhu' | /CV-C [´] V [˘] C-CV-CV-CV/ | 'his teacher' |

- When a word contains two or more long syllables, the long syllable nearest to the end of the word (the very last syllable does not count) receives the primary stress, and in most cases the one closest to the beginning receives the secondary stress.

| | | |
|--------------------|---------------------------|--|
| 'raʔiisuhúnna' | /CV-CV̄V-CV-CV̄C-CV/ | (their chief) (for feminine plural) |
| 'mustāwdaēāatuhum' | /CVC-CV̄C-CV-CV̄V-CV-CVC/ | (their deposits) (for masculine plural) |

3.9.2 Duration

The duration of a sound is the actual time taken to produce it. The relative duration of certain phonemes depends on several factors such as speaking rate, speaking style (reading versus conversation), stress, the location of pauses and of word and syllable boundaries, the place and manner of articulation, and the rhythm.

Duration is significant in the Arabic language and a difference in the length of a vowel or consonant, makes a difference in meaning.

The three short vowels in Arabic may be prolonged to their long counterparts. The difference between short and long vowels is approximately double or more. Increasing the vowel duration may change the meaning of the carrier word. For example:

| | | |
|--------------------|-------------|---------------------------------------|
| 'sin' | /CVC/ | (tooth) |
| 'siin' | /CVVC/ | (the letter 's') |
| 'kataba' | /CV-CV-CV/ | (he wrote) |
| 'kaataba' | /CVV-CV-CV/ | (he exchanges letters with some body) |
| ' <u>k</u> atala' | /CV-CV-CV/ | (he killed) |
| ' <u>k</u> aatala' | /CVV-CV-CV/ | (he fought) |

The duration of the short vowels is from 100-160 msec, with the long vowels it is from 200-350 msec.

The duration of consonants depends upon whether they occur initially, medially or finally. It also depends on other conditions, namely on the manner of articulation, voiced

or unvoiced, and single or geminated. The duration of consonants varies from 40-375 msec [103].

3.9.3 Intonation

In Arabic as well as in many European languages, intonation is used to convey grammatical form. The acoustical correlate of the intonation contour is the fundamental frequency (F0) of the vocal cords (excitation source), as a function of time. The pitch or F0 ranges between 80-160 Hz for male speakers, and between 160-400 Hz for female speakers, therefore the changes in F0 are more important than its absolute values.

F0 trends (directions of changes) may change several times in a single phone and thus may signal stress or syntactic information via both its relative values and its slopes. Most phones have a simple rising or falling F0 pattern, but a single phone may contain a rise-fall-rise contour. As a result, there are global changes in F0 over a sentence and local changes over a phone. The shapes of the intonation pattern depends on the sentence type (i.e., declaration, commands, question, etc.) [114].

3.10 Summary

The Arabic language has 6 vowels and 29 consonants. The Arabic consonantal system differs from the Latin one, primarily to the presence of the pharyngealised and glottal phonemes. The consonants can be classified mainly into 11 broad phonetic classes. The Arabic language uses only three syllabic types (i.e., /CV/, /CVC/, and /CVCC/).

The statistical results show that the six vowels represent about 43% of the total number of phonemes, while the 29 consonants represent 57%, (using 75875 phonemes). The vowel /a/ has the highest percentage (47%) among the six vowels. The results also show that the syllable /CV/ represents about 67% of the 75875 checked syllables.

Arabic words are categorised according to their morphological patterns (balances). The balance shows the syllabic structure and the actual vowels of a word. The language uses a limited number of balances to describe verbs, while it uses a large number of balances for nouns. These balances can be considered as a vital tool in Arabic speech recognition, where the morphological pattern of a word gives grammatical, syntactic, and semantic information about that word.

Chapter 4

A Model of Lexical Access for a Large Vocabulary Speech Recognition System

4.1 Introduction

Isolated word speech recognition can be achieved using either a mathematical or an acoustic-phonetic approach.

In the mathematical approach, pattern matching and stochastic modelling using hidden Markov models are used. An unknown word is matched against all vocabulary reference patterns or models (using whole word templates or models). Generally speaking, these methods utilise little or no speech-specific knowledge, and the storage requirement and computation both increase almost linearly with the vocabulary size. The introduction of some speech compression techniques such as vector quantisation, has led to drastic cuts in storage and computation (see chapter 1).

The extension of these techniques to multiple speakers, large vocabularies, and /or continuous speech is highly questionable. Even if the computational and storage costs are not an issue, the drawback of these techniques is the amount of work that must be done for the creation of the reference templates and their updating for speaker-independent systems. Also, the performance of these techniques would surely deteriorate for a large vocabulary system, mainly due to the increase in the probability of the existence of acoustically similar words.

A suitable alternative to the mathematical approach, for recognition of utterances from large vocabularies, is the use of the acoustic-phonetic approach. In this approach, the speech acoustic signal is mapped (segmented and labelled) into a sequence of linguistic units such as phonemes, diphones, demisyllables, or syllables. The resulting string of units (labels) is used for lexical and syntactic analysis. Words in the lexicon are represented in terms of phonemic spellings.

The major problem with this method is our inability to extract phonetic information reliably from the speech signal due to the variability in the acoustic realisation of utterances. This variability can come from diverse sources, such as the speaking environment, the position and characteristics of the transducer, and inter-speaker variabilities. The latter variabilities can result from changes in the speaking rate, differences in voice quality according to the speaker's physiological and psychological state, and differences across-speakers (i.e., in vocal tract size and shape, sociolinguistic background, and dialect).

The acoustic realisation of the speech signal conveys linguistics and extra linguistic information (e.g., acoustic environment, identity of the speaker, his physiological and psychological states, etc.). Successful speech recognition is possible only if we can extract the linguistic information while discarding other irrelevant information. An alternative to detailed acoustic-phonetic analysis is the use of broad phonetic analysis.

4.2 Broad Phonetic Classification and Lexical Access

One way of discriminating between words in a large vocabulary is through the use of broad phonetic classes. In this case, the speech signal is segmented by coarse reliable acoustic analysis in terms of broad phonetic classes [115, 116, 117]. The broad phonetic description of a word is used as a means of lexical access. The lexicon is structured into sets of words sharing the same broad phonetic labelling, which are called 'cohorts'. In this way, a substantial reduction in the number of possible word candidates which match an unknown word, can be obtained for a large vocabulary recognition system.

Figure 4.1 shows a block diagram of a word recognition system based on the broad phonetic approach. This model includes three stages, namely classification of the acoustic signal, lexical access, and verification. The classification and lexical access can be done in a bottom-up phase and the verification in a top-down phase. In the bottom-up phase, the sequence of broad phonetic classes is extracted from the acoustical signal and used to retrieve a set of word candidates from a large lexicon. In the top-down phase, the constraints imposed by the phonemic structure of the chosen set of words, select and schedule the verification process. In this process, context-dependent procedures which are most appropriate for performing detailed phoneme verification analyses, in delimited signal intervals, are used to determine among the word candidates the most likely spoken one.

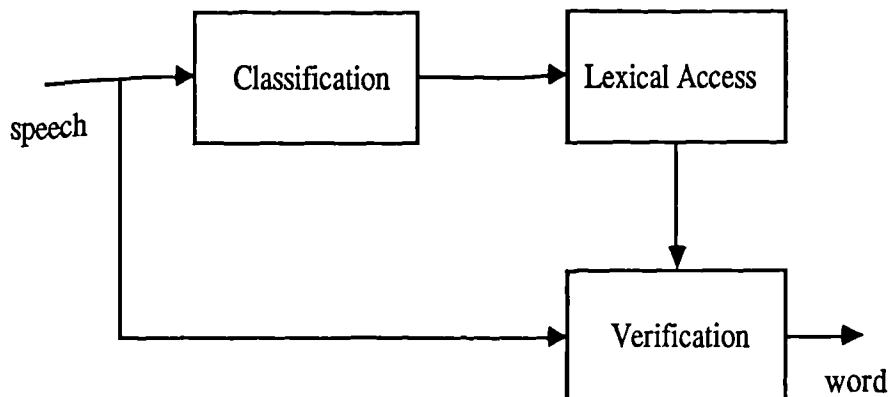


Fig. 4.1 Word recognition system

This recognition model has two major features. First, the classification of the speech signal is in terms of broad phonetic units instead of linguistic ones. Second, there is no attempt to perform detailed recognition of the acoustic signal until after lexical access, when phonetic context can be used to aid the recognition. That is, detailed analysis is only performed in answering specific recognition hypotheses during verification.

The choice of specific broad phonetic classes greatly affects the performance of the recognition system. If very broad classes are extracted from the acoustic signal, then the error rate in recognising these classes will be very low. However, a large number of words will match any sequence of the broad classes. If the classes are detailed, the error rate will be higher, but fewer words will match any sequence. Thus, the problem is one of finding a representation which is broad enough to have a low error rate and narrow enough for a class sequence to match a small number of words.

The representation of broad phonetic classes (BPCs) is based on manner of articulation classes [115], where manner-of-articulation differences tend to be relatively invariant across different speakers and different phonetic contexts.

A number of investigations on the consequences of using broad phonetic classification have been conducted for different languages, e.g. for English, Dutch, and Italian. These studies focus on the influence of different choices of BPC on the size and number of

'cohorts'. By using six BPCs, namely vowel, plosive, nasal, liquid, strong fricative, and weak fricative, with a database consisting of 20,000 words of American English, it was found that the maximum cohort size was 223 words (about 1% of the lexicon size), and almost one-third (32%) of the words were uniquely represented at this broad phonetic level [115]. In another study (on Dutch), five BPCs were used to classify a lexicon of 11644 words. It was found, that the maximum cohort size was 16, and about 12% of the words were uniquely represented at this level [118]. Also, using a lexicon of 12266 Italian words (roots), the maximum cohort size was 15 and 68% of the words were uniquely represented by employing nine BPCs [119]. In the following sections, the results of performing several classification schemes on the Arabic language are reported.

These results show that the sound patterns of a given language are limited not by only the inventory of basic sound units, but also by the allowable combinations of these sound units.

4.3 Discrimination of Words

Broad phonetic classification has been applied to large vocabulary lexicons. The lexical database, which contains two lexicons (the phonemic form of 3,000 words, and 10,000 words, see chapter 1), has been used.

Our aim in this study is to try as much as possible to use a set of BPCs which can identify most (or all) of the words uniquely with a minimum of detailed acoustic information. Such a set is going to be an optimal interaction between the properties of the lexicon and the possibilities of acoustic analysis.

In the previous chapter, we have seen the importance of the morphological pattern or balance in the Arabic language. This balance describes basically the syllabic structure and the actual vowels used in a word. For example, take the case of monosyllabic words, where we may have up to 15 different balances as illustrated in the diagram of Figure 4.2. In this diagram, there are 6 branches at the first level which correspond to the six vowels. Each short vowel may fall into one of three different syllabic patterns, and each long vowel may fall into one of two different syllabic patterns, resulting in a total of 15 possible balances for the monosyllabic words.

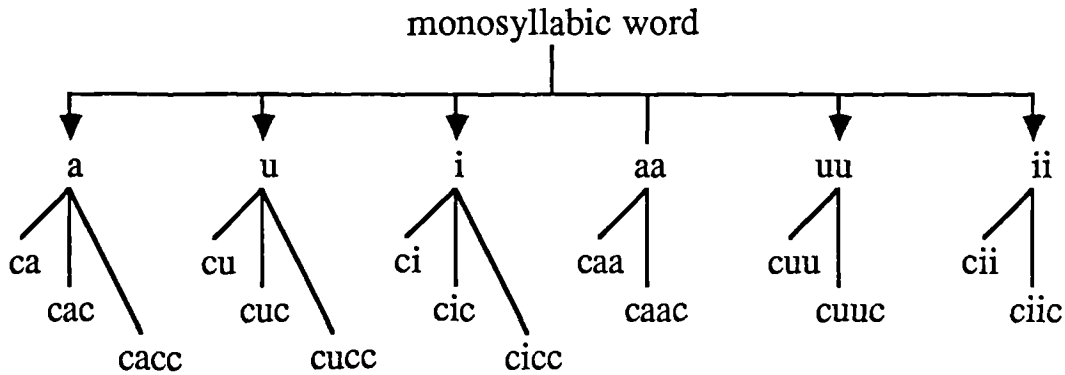


Fig. 4.2 Possible balances for monosyllabic words

On the same basis, for disyllabic words we may have up to 180 possible balances, where the second syllable has 15 possible patterns, as in the case of monosyllabic word and the first syllable can take only 12 possible patterns (the pattern /CVCC/ does not occur in word-initial position). For n-syllable words, we may have :

- $(6)^n$ possible vowel combinations (branches) at the first level
- $15 \cdot (12)^{n-1}$ possible balances at the second level

From the above demonstration, we notice that by just looking at the sequence of vowels and the syllabic structure of a word, we can have a large number of possible balances and hence a large number of possible 'cohorts'.

Identifying the morphological balance in a bottom-up procedure in an Arabic recognition system, facilitates the use of linguistic knowledge in a top-down recognition phase, where a morphological analyser based on the root-balance, can identify the syntactic structure of a word and give an indication of its meaning. The morphological balance is therefore a key factor in an Arabic speech recognition system.

The Arabic language uses hundreds of morphological balances, (where Arabic words are already categorised according to morphological balances). The verbs can take just a few hundred balances (about 400 balances are used for describing the verbs [122]), and

nouns take a large number of balances [120, 121]. By considering the most frequent balances, it was found that less than 1000 patterns (balances) are enough to be used in a morphological compression procedure for Arabic text [123].

The consonants of each morphological balance can be classified according to broad phonetic classes. Different classification schemes are given in the next section.

4.3.1 Classification Schemes

Ten different classification schemes are considered in this study [124]. These schemes can be divided into two similar groups. In the first group, vowels are replaced with the symbol /V/, while in the second group vowels retain their phonetic symbol (i.e., /a/, /u/, /i/, /aa/, /uu/, and /ii/). These schemes are as follows:

First group: in this group vowels are replaced by the symbol /V/

- 1) C/V Consonants are replaced by the symbol /C/.
- 2) 4BPC/V Consonants are classified according to four BPC: voiced plosive, unvoiced plosive, unvoiced fricative, and other voiced consonants.
- 3) 5BPC/V Consonants are classified according to five BPC: plosive, fricative, nasal, liquid, and semivowel.
- 4) 7BPC/V Consonants are classified according to seven BPC: voiced plosive, unvoiced plosive, voiced fricative, unvoiced fricative, nasal, liquid, and semivowel.
- 5) 11BPC/V Consonants are classified according to eleven BPC: pharyngealised voiced plosive, non-pharyngealised voiced plosive, pharyngealised unvoiced plosive, non-pharyngealised unvoiced plosive, pharyngealised voiced fricative, non-pharyngealised voiced fricative, pharyngealised unvoiced fricative, non-pharyngealised unvoiced fricative, nasal, liquid, and semivowel. Details of these BPC are given in Table 3.3 .

Second group: in this group vowels are classified according to one of the six Arabic vowels, and consonants are classified as in the first group, giving the following schemes:

- 6) C/6V
- 7) 4BPC/6V
- 8) 5BPC/6V
- 9) 7BPC/6V
- 10) 11BPC/6V

The sixth classification scheme (in the second group C/6V), represents a classification according to morphological balances.

4.3.2 Statistical Results

The results of using the ten classification schemes are summarised in tables 4.1 and 4.2. Table 4.1 shows the results for the 3000-word lexicon, while Table 4.2 gives the results for the 10,000-word lexicon.

From Table 4.1, we notice that on one hand, the number of unique word cohorts (i.e., cohorts having just one word), increases with the number of BPCs, while on the other hand, the maximum cohort size (maximum number of words in a cohort), decreases as the number of BPCs increases. For the C/V classification scheme, the 3000 words are grouped in just 31 cohorts, while using the C/6V scheme they are grouped in 286 cohorts (morphological balances). The percentage of uniquely represented words rises from about 55% for 11BPC/V scheme to about 82% for the 11BPC/6V scheme. In the latter case, the maximum cohort size is 5 and the average cohort size is just 1.11.

The classification results of the second lexicon (10,000 words) given in Table 4.2, are almost similar to that of the first lexicon. However in this table, there is a rise in the percentage of unique word cohorts for all the classification schemes compared with that of Table 4.1. This is mainly due to the increase in the number of polysyllabic words in the second lexicon.

The percentage of uniquely represented words has also risen here from about 48% when using 11BPC/V scheme, to about 89% when employing the 11BPC/6V. Even for simple classification scheme (i.e., comparable with other schemes) such as the 4BPC/6V, the percentage of uniquely represented words (about 54%) is high, the maximum cohort size is 17 words, and the average cohort size is 1.44 word.

| | Vowel | | | | | 6 Vowels | | | | |
|---------------------------------|-------|------|-------|-------|-------|----------|-------|-------|-------|-------|
| | C | 4BPC | 5BPC | 7BPC | 11BPC | C | 4BPC | 5BPC | 7BPC | 11BPC |
| no. of cohorts | 31 | 868 | 1079 | 1835 | 2156 | 286 | 1762 | 1972 | 2558 | 2714 |
| no. of unique word cohorts | 3 | 420 | 527 | 1244 | 1651 | 134 | 1210 | 1432 | 2226 | 2479 |
| maximum cohort size | 599 | 66 | 39 | 19 | 15 | 174 | 16 | 13 | 6 | 5 |
| average cohort size | 96.77 | 3.46 | 2.78 | 1.63 | 1.39 | 10.49 | 1.70 | 1.52 | 1.17 | 1.11 |
| % of uniquely represented words | 0.1 | 14. | 17.56 | 41.46 | 55.03 | 4.46 | 40.33 | 47.73 | 74.20 | 82.63 |

Table 4.1 Classification results for the 3000-word lexicon

| | Vowel | | | | | 6 Vowels | | | | |
|---------------------------------|-------|-------|-------|-------|-------|----------|-------|-------|-------|-------|
| | C | 4BPC | 5BPC | 7BPC | 11BPC | C | 4BPC | 5BPC | 7BPC | 11BPC |
| no. of cohorts | 72 | 2981 | 3722 | 5810 | 6654 | 1437 | 6922 | 7579 | 9043 | 9384 |
| no. of unique word cohorts | 9 | 1518 | 2048 | 3862 | 4785 | 683 | 5365 | 6180 | 8317 | 8876 |
| maximum cohort size | 1022 | 89 | 54 | 28 | 22 | 197 | 17 | 15 | 6 | 5 |
| average cohort size | 138.8 | 3.35 | 2.69 | 1.72 | 1.50 | 6.96 | 1.44 | 1.32 | 1.11 | 1.07 |
| % of uniquely represented words | 0.09 | 15.18 | 20.48 | 38.62 | 47.85 | 6.83 | 53.65 | 61.80 | 83.17 | 88.76 |

Table 4.2 Classification results for the 10,000-word lexicon.

In general, the detailed vowel classification has almost doubled the number of uniquely represented words (e.g., from 38% for the 7BPC/V scheme to 83% for the 7BPC/6V scheme), and has led also to specifying the morphological balance of a word.

4.3.3 Effect of Syllabic Structure on the Classification Results

Words in the two lexicons vary in their size from 1 to 7 syllables as follows:

| lexicon | no of syllables in a word | | | | | | |
|---------|---------------------------|------|------|------|-----|-----|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 191 | 1505 | 1077 | 217 | 10 | - | - |
| 2 | 193 | 2096 | 3971 | 2624 | 924 | 188 | 4 |

Tables 4.3 to 4.8 Show the classification results with respect to the the number of syllables in a word for the second lexicon.

Table 4.3 displays the results for monosyllabic words (193 words). It shows that the percentage of uniquely represented words goes up to 77% of the total number of words for the 11BPC/6V scheme. None of the words is uniquely represented using the C/V scheme.

Table 4.4 displays the results on disyllabic words (2096 words). It shows that the percentage of uniquely represented words can reach up to 80% of the total number of words for the 11BPC/6V scheme, which is higher than that of monosyllabic words.

Table 4.5 shows the results for trisyllabic words (3971 words). It shows that the percentage of uniquely represented words reaches about 87% of the total number of words for the 11BPC/6V scheme. Only 2.64% of the total words are uniquely represented by the C/6V scheme (balances). Trisyllabic words are the most common words in the language.

| | Vowel | | | | | 6 Vowels | | | | |
|---------------------------------|-------|------|-------|-------|-------|----------|-------|-------|-------|-------|
| | C | 4BPC | 5BPC | 7BPC | 11BPC | C | 4BPC | 5BPC | 7BPC | 11BPC |
| no. of cohorts | 3 | 53 | 75 | 122 | 136 | 13 | 102 | 128 | 162 | 168 |
| no. of unique word cohorts | 0 | 18 | 27 | 75 | 95 | 3 | 66 | 91 | 140 | 150 |
| maximum cohort size | 123 | 18 | 10 | 8 | 7 | 84 | 15 | 9 | 4 | 4 |
| % of uniquely represented words | 0 | 9.32 | 13.98 | 38.86 | 49.22 | 1.55 | 34.19 | 47.15 | 72.53 | 77.72 |

Table 4.3 Classification results for 193 monosyllabic words

| | Vowel | | | | | 6 Vowels | | | | |
|---------------------------------|-------|------|------|-------|-------|----------|-------|-------|-------|-------|
| | C | 4BPC | 5BPC | 7BPC | 11BPC | C | 4BPC | 5BPC | 7BPC | 11BPC |
| no. of cohorts | 5 | 283 | 427 | 913 | 1168 | 89 | 1069 | 1288 | 1726 | 1866 |
| no. of unique word cohorts | 1 | 67 | 121 | 454 | 698 | 15 | 629 | 878 | 1446 | 1680 |
| maximum cohort size | 859 | 80 | 49 | 19 | 15 | 192 | 16 | 13 | 6 | 5 |
| % of uniquely represented words | 0.04 | 3.19 | 5.77 | 21.66 | 33.03 | 0.07 | 30.00 | 41.88 | 68.98 | 80.15 |

Table 4.4 Classification results for 2096 disyllabic words

| | Vowel | | | | | 6 Vowels | | | | |
|---------------------------------|-------|------|-------|-------|-------|----------|-------|-------|-------|-------|
| | C | 4BPC | 5BPC | 7BPC | 11BPC | C | 4BPC | 5BPC | 7BPC | 11BPC |
| no. of cohorts | 9 | 851 | 1085 | 2008 | 2425 | 338 | 2637 | 2865 | 3544 | 3687 |
| no. of unique word cohorts | 1 | 326 | 471 | 1207 | 1651 | 105 | 1966 | 2232 | 3234 | 3454 |
| maximum cohort size | 1022 | 89 | 54 | 28 | 22 | 197 | 17 | 15 | 5 | 4 |
| % of uniquely represented words | 0.02 | 8.20 | 11.86 | 30.39 | 41.57 | 2.64 | 49.50 | 56.20 | 81.44 | 86.98 |

Table 4.5 Classification results for 3971 trisyllabic words

| | Vowel | | | | | 6 Vowels | | | | |
|---------------------------------|-------|-------|-------|-------|-------|----------|-------|-------|-------|-------|
| | C | 4BPC | 5BPC | 7BPC | 11BPC | C | 4BPC | 5BPC | 7BPC | 11BPC |
| no. of cohorts | 16 | 1055 | 1292 | 1830 | 1976 | 540 | 2109 | 2253 | 2509 | 2556 |
| no. of unique word cohorts | 1 | 569 | 770 | 1333 | 1535 | 255 | 1786 | 1991 | 2408 | 2494 |
| maximum cohort size | 759 | 33 | 15 | 10 | 9 | 124 | 9 | 6 | 4 | 4 |
| % of uniquely represented words | 0.03 | 21.68 | 29.34 | 50.80 | 58.49 | 9.71 | 68.06 | 75.87 | 91.76 | 95.04 |

Table 4.6 Classification results for 2624 quadrisyllabic words

| | Vowel | | | | | 6 Vowels | | | | |
|---------------------------------|-------|-------|-------|-------|-------|----------|-------|-------|-------|-------|
| | C | 4BPC | 5BPC | 7BPC | 11BPC | C | 4BPC | 5BPC | 7BPC | 11BPC |
| no. of cohorts | 21 | 578 | 670 | 759 | 771 | 350 | 817 | 858 | 910 | 915 |
| no. of unique word cohorts | 4 | 401 | 503 | 628 | 641 | 232 | 734 | 806 | 897 | 906 |
| maximum cohort size | 262 | 14 | 7 | 5 | 4 | 48 | 6 | 4 | 3 | 2 |
| % of uniquely represented words | 0.4 | 43.39 | 54.43 | 67.96 | 69.37 | 25.10 | 79.43 | 87.22 | 97.07 | 98.05 |

Table 4.7 Classification results for 924 five-syllable words

| | Vowel | | | | | 6 Vowels | | | | |
|---------------------------------|-------|-------|-------|-------|-------|----------|-------|-------|------|-------|
| | C | 4BPC | 5BPC | 7BPC | 11BPC | C | 4BPC | 5BPC | 7BPC | 11BPC |
| no. of cohorts | 17 | 157 | 169 | 174 | 174 | 106 | 184 | 183 | 188 | 188 |
| no. of unique word cohorts | 2 | 133 | 152 | 161 | 161 | 73 | 180 | 178 | 188 | 188 |
| maximum cohort size | 45 | 4 | 3 | 3 | 3 | 11 | 2 | 2 | 1 | 1 |
| % of uniquely represented words | 1.06 | 70.74 | 80.85 | 85.63 | 85.63 | 38.82 | 95.75 | 97.34 | 100 | 100 |

Table 4.8 Classification results for 188 six-syllable words

Table 4.6 displays the results for quadrisyllabic words (2624 words). It shows that the percentage of uniquely represented words reaches about 95% of the total number of words for the 11BPC/6V scheme. 9.7% of the words are uniquely represented using the C/6V scheme.

Tables 4.7 and 4.8 show the classification results for the five-syllable words (924 words), and for the six-syllable words (188 words), respectively. The percentages of uniquely represented words are about 98% for the five-syllable words, and 100% for the six syllable words, using the 11BPC/6V scheme.

For seven-syllable words (only 4 words), all are uniquely represented by all the schemes except the two schemes C/V and C/6V.

As a result, the increase in the number of syllables in a word has led to a higher percentage of unique word cohorts. Polysyllabic words are more likely to be classified in a unique word cohort, because of the limitation imposed by the number of allowable or used combinations in the language itself (phonological limitations).

4.3.4 Discussion

The use of the above mentioned classification schemes leads to drastic cuts in the number of word-candidates at the lexical level for a specific pattern. Using broad phonetic classification for consonants and detailed vowel classification has led to a powerful lexical access for the Arabic language. It has also given at the same time some information about the morphological balance of a word, which is very important for higher level sources of knowledge, especially in continuous speech recognition or speech understanding systems.

These findings about phonological constraints have important implications for speech recognition. They suggest that a complete and detailed phonetic analysis of the speech signal is not only undesirable from an error propagation standpoint, but may indeed be unnecessary.

We have seen that the maximum cohort size is 5 for the 11BPC/6V scheme. Prosodic information such as stress position, and duration of different phonetic segments could also be very important potential cues for reducing the cohort size.

4.4 Structure of the Lexicon

In a recognition system based on broad phonetic analysis, words are usually stored in the lexicon into tables according to their broad phonetic description (sequence of labels). Therefore a simple table look-up produces the set of words matching a given sequence of broad phonetic classes.

For the Arabic language, we suggest that the lexicon is structured into a hierarchical form (tree form) as follows:

- Number of syllables in a word
 - Sequence of Vowels
 - Morphological balance (syllabic structure)
 - Hierarchical broad phonetic description for consonants, where the last branch contains the set of words sharing the same labelling.

In this hierarchical arrangement, words can easily be looked-up at a given phonetic description. This description can begin with a very simple one such as the number of syllables in a word, and can end with the broad phonetic descriptions of consonants and actual vowels. The hierarchical broad phonetic description for consonants is illustrated in Figure 4.3.

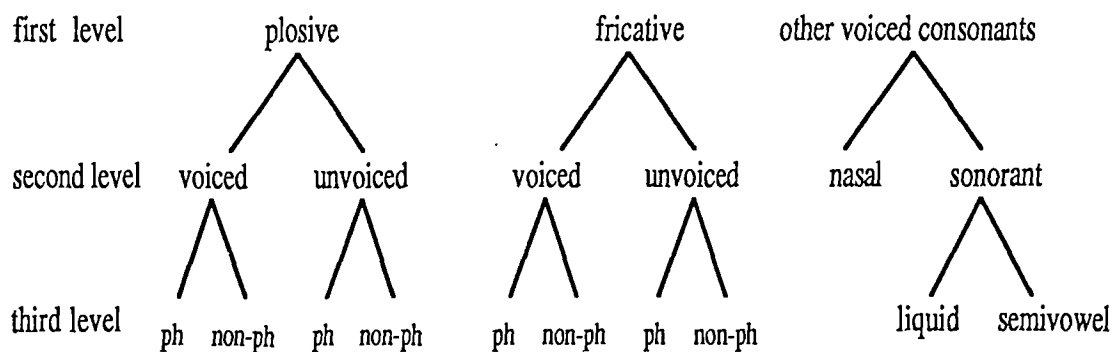


Fig. 4.3 Hierarchical description of consonants
(ph: refers to pharyngealised)

According to the above lexical structure, different speech recognition systems (i.e., small, medium, and large vocabulary systems) can use this lexicon.

The proposed structure and organisation of the lexicon are designed for speeding up the retrieval of a small sub-set of words. In some cases, where the acoustic realisations of some consonants are ambiguous (or some acoustical events are deleted or missed), one can stay at an upper level in this tree and employ extra linguistic knowledge to assist in recovering the missing information (in a continuous speech recognition system).

4.5 An Automatic Speech Recognition Model

The proposed model for a large vocabulary isolated word speech recognition system is given in Figure 4.4. In this model the speech signal is first transformed into acoustic parameters through the feature measurement stage. The parameter complexity depends on the employed set of BPCs in the broad phonetic segmentation stage. These parameters are used in the vowel recognition stage and in the segmentation stage. The output of the vowel recognition stage is fed to the segmentation stage, which gives at its output a sequence of phonetic labels which are used for lexical access (bottom-up phase). The result of the lexical access is a small set of word candidates (or more likely a single word candidate), sharing the same phonetic labelling. In the last stage of this model, differential diagnostic techniques are used, in conjunction with detailed acoustic cues from the speech signal, to select the most likely word candidate (top-down phase).

This model relies on the fact that the constraints imposed by the language on possible sound patterns should significantly reduce the number of word candidates. It is also computationally efficient, since detailed acoustic knowledge is applied in a top-down verification mode just when it is needed.

In Section 4.3.1 different classification schemes have been presented. A recognition system based on a hierarchical classification can start with detailed vowel recognition and a simple set of broad phonetic classes. The resultant sequence of labels is used for lexical access. If the number of word candidates is more than one, the system goes back to the segmentation process and widens the set of BPCs until it reaches the minimum possible number of word candidates. Then the verifier starts its process, if the number of word candidates exceeds one.

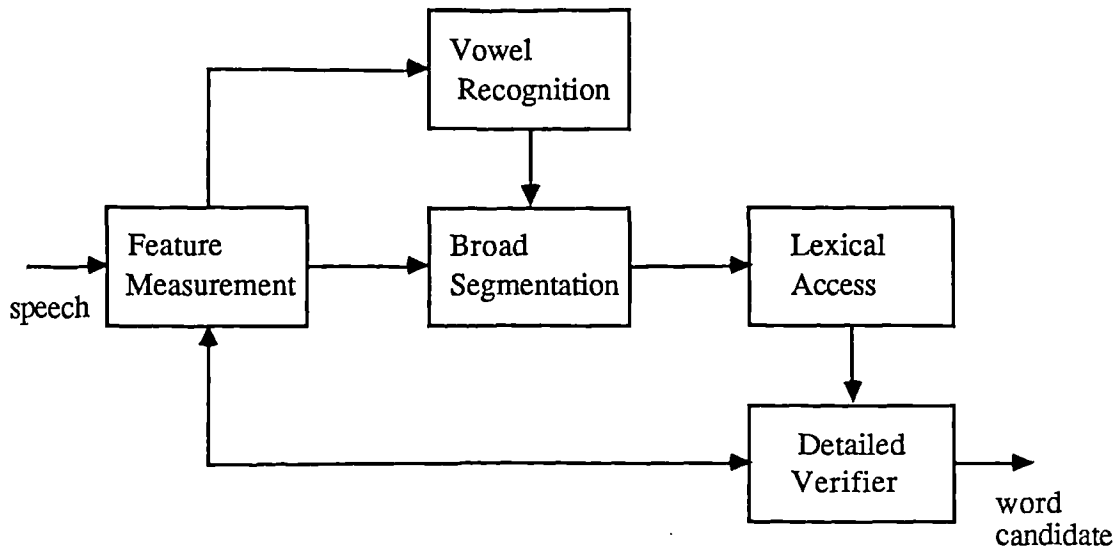


Fig. 4.4 A speech recognition model

4.6 Speech Recognition Experiment

An experiment has been carried out to investigate the possibility of performing broad phonetic segmentation for consonants and detailed vowel recognition according to the model of Figure 4.4. In this experiment, a basic speech database consisting of one hundred words uttered by three male speakers has been used. Some words in this set uttered by different speakers (i.e., male and female), have also been tested.

The processes of the bottom-up phase of Figure 4.4 are addressed in the following chapters of this thesis. The verification stage is not addressed in this thesis.

Figure 4.5 shows the processes performed in the bottom-up phase of the speech recognition model considered in this research work. Five main procedures have been developed during this phase, i.e.:

- Voiced-unvoiced-silence segmentation (V-UV-S).
- Vowel recognition.
- Spectral variation (transition) detection.
- Broad phonetic segmentation procedure (according to scheme number 7 given in

Section 4.3.1).

- Error correction procedure.

The first two procedures are addressed in chapter 5. The third one is introduced in chapter 6, while the fourth and fifth procedures are demonstrated in chapter 7. The last chapter summarises the whole research work starting from the classification scheme passing through the above five procedures and ending with a discussion and some suggestions for further research.

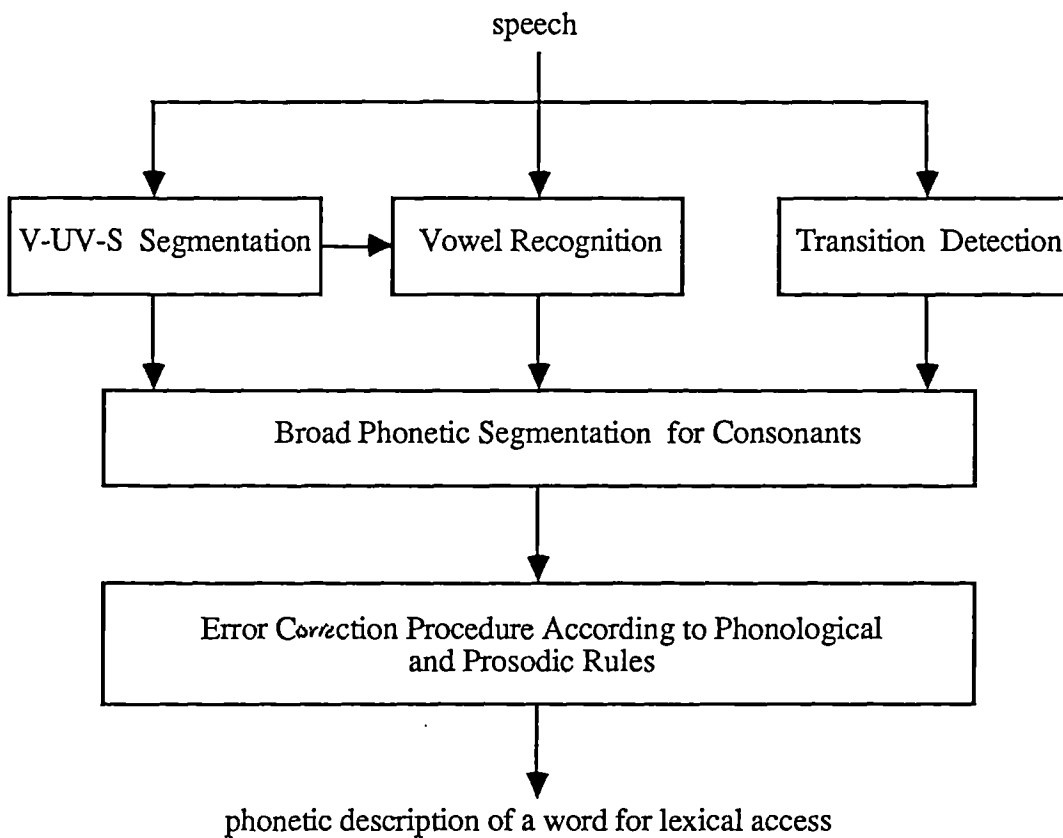


Fig. 4.5 Processes of the bottom-up phase

4.7 Summary

Word discrimination according to broad phonetic classes has been presented in this chapter. Different sets of classification schemes are used to describe large vocabulary lexicons. The phonetic description in these schemes varies from 2 to 17 phonetic classes.

The statistical results show that about 89% of the 10,000 tested words can be uniquely represented by using 11 broad phonetic classes for consonants and six classes for vowels. In this case, the maximum number of words having the same phonetic labelling is 5.

A word recognition model has been proposed in this chapter. This model performs detailed vowel recognition and broad phonetic analysis for consonants. The resultant string of labels of an unknown input word is used for lexical access to retrieve the word (or the set of words) having the same phonetic description of the input word. Finally, the verifier in this recognition model is activated if the number of word candidates exceeds one. The verification stage is not addressed in this thesis. The vowel recognition, segmentation and labelling processes are described in the following chapters.

Chapter 5

Preliminary Segmentation and Vowel Recognition

5.1 Introduction

Two stages from the speech recognition system of Figure 4.5, namely V-UV-S segmentation and vowel recognition procedures, are demonstrated in this chapter (see Figure 5.1).

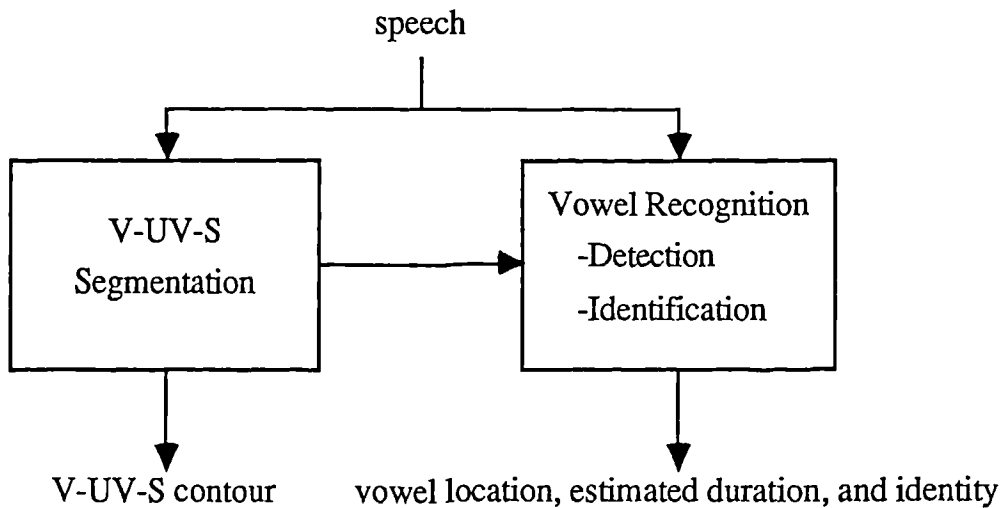


Fig. 5.1 Preliminary V-UV-S segmentation and vowel recognition

Preliminary segmentation is carried out on the speech signal, where a given speech frame is classified as voiced speech (V), unvoiced speech (UV), or silence (S) (absence of speech). A string of adjacent frames sharing the same labelling are grouped together and called a voiced segment, an unvoiced segment, or a silence segment. A special smoothing

algorithm is applied on the resultant decision to remove any spurious frame (or frames) between different segments. The outcome of this preliminary segmentation is the V-UV-S contour which is employed in the segmentation procedure presented in Chapter 7. The detection of word boundaries (endpoints) is performed through the voiced-unvoiced-silence detection process.

The aim of the vowel recognition procedure is to recognise the six Arabic vowels, i.e., the three short vowels /a/, /u/, and /i/, and their three long counterparts /aa/, /uu/, /ii/. This procedure consists of two main phases, namely vowel detection and vowel identification. Vowel locations are spotted within voiced regions in the vowel detection phase, while vowel identities are determined in the identification phase.

5.2 Voiced-Unvoiced-Silence Segmentation

A variety of approaches have been described in the speech literature for making the voiced-unvoiced-silence (V-UV-S) decision [125-126]. The complexity of the voiced-unvoiced algorithm depends on the bandwidth of the input speech signal. For a wide band speech signal (up to 8 KHz), this decision could be easily measured for a short-time speech signal (frame) by taking the ratio of the signal energy below 1 KHz to that above 5 KHz. If this ratio exceeds a certain level, the frame is tagged as a voiced frame, otherwise it is tagged as an unvoiced frame [127]. For a telephone speech signal (up to 3.4 KHz), several parameters such as energy, zero-crossing rate, some LPC coefficients, autocorrelation coefficients, etc. [126], were employed for the V-UV-S decision. In our system, several algorithms have been tested, and the adopted algorithm performs V-UV-S detection in conjunction with pitch analysis.

The speech signal (in the speech database) is lowpass filtered to 4800 Hz, sampled at 10 KHz, and each sample is quantised with an accuracy of 12 bits. Then, the speech signal is highpass filtered at 60 Hz to remove any dc, low-frequency hum, or noise components which might be present in the speech signal. The resultant speech signal is grouped into blocks of size $N=256$ samples (25.6 msec) for the pitch computation to allow for at least two pitch periods to be present in one block. For the computation of other parameters such as energy and zero-crossing rate, a block of 128 samples (12.8 msec) is used. All parameters are computed at every 6.4 msec (i.e., about 156 times per second), with 50% overlap for energy blocks and 75% for pitch blocks.

The following measurements have been used in the V-UV-S segmentation procedure:

- 1) Energy of the signal $s(n)$ in the range 60-4800 Hz.
- 2) Energy of the signal $s_h(n)$ in the range 300-4800 Hz.
- 3) Normalised autocorrelation coefficient at unit sample delay.
- 4) Zero-crossing rate of the signal.
- 5) Pitch value.

Figure 5.2 shows a block diagram of the processes involved in the V-UV-S decision. These processes are explained in the following sections.

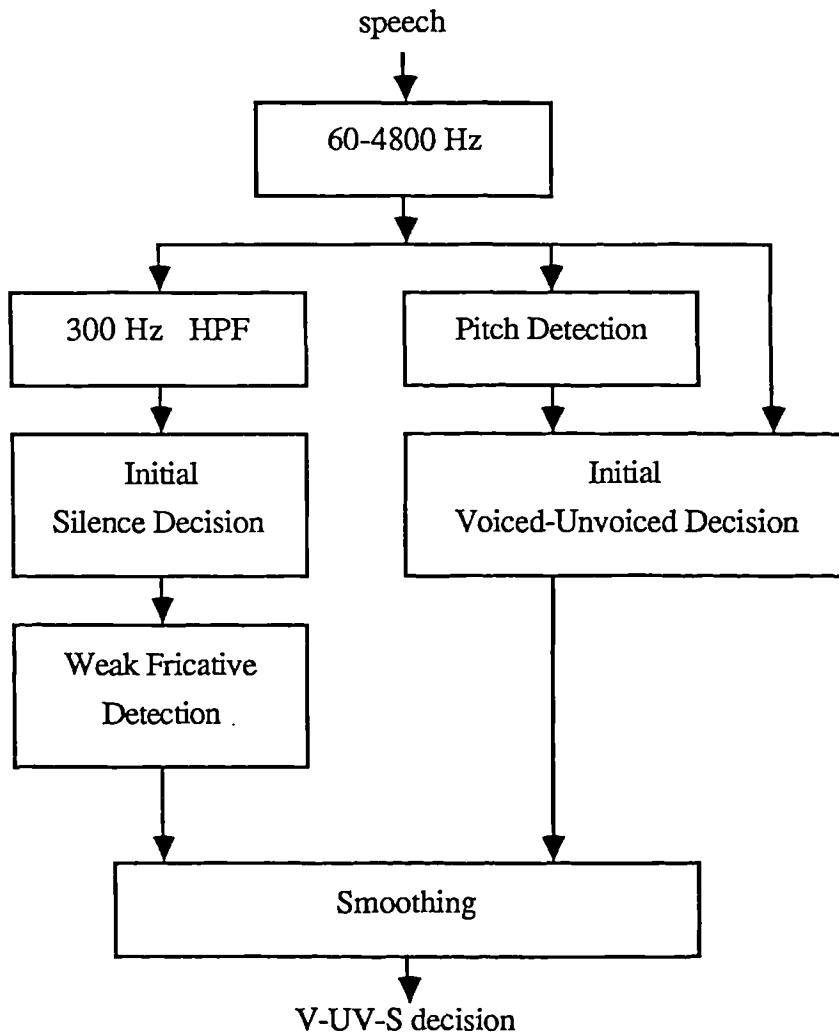


Fig. 5.2 Block diagram of the V-UV-S decision

5.2.1 Initial Silence Detection

Initial silence detection is based on the momentary speech signal energy. Energy is defined as:

$$E_h = \frac{1}{N} \sum_{n=1}^N s_h^2(n) \quad (5.1)$$

where N is the number of samples in the analysis block. The $s_h(n)$ is the digitised speech signal, where the subscript h refers to the highpass filtered version in the range 300-4800 Hz. The reason for employing the energy above 300 Hz in detecting the silence is to allow voiced plosive sounds, which might have a low frequency voice bar, to be detected as silence. If the pitch detector indicates the presence of voicing throughout the duration of a silence segment, the segment is tagged as a voiced silence, or voice bar as will be explained in Chapter 7, in the plosive detection section.

If E_h falls below a certain threshold, the speech frame (6.4 msec) is tagged as a silence frame. This threshold is called the silence threshold (S_{th}), and it is chosen to be about 6 dB higher than the average value of the background noise level on the input analogue tape (where the original speech was recorded). In our case S_{th} is taken equal to 400 (linear scale). The background noise can be measured for each word (or for the entire recording session), where there is no speech in the first few msec of the recording interval before the beginning of each word. In our case, a noise cancelling microphone has been used, where the background noise is very low and actually equals the noise introduced by the tape recorder.

The endpoints of a word are automatically detected by locating the silence segments at the word boundaries. In general, about 200 msec of silence before the beginning and after the end of a word are stored for voiced plosive detection.

5.2.2 Pitch Detection

Because of the importance of pitch detection (e.g., for vocoder, for speaker recognition, etc.), a wide variety of algorithms for pitch detection have been proposed in the speech processing literature. A comparative performance study of some pitch detection

algorithms is given in [128]. Two pitch detection algorithms have been implemented and tested in this study, namely the autocorrelation method [129], and the cepstral method [130]. A pilot experiment has shown more reliable results by the former method than by the latter. For this reason, the autocorrelation method has been maintained for all later experiments.

Figure 5.3 shows a block diagram of the autocorrelation pitch detector. In this method the speech signal is lowpass filtered to 900 Hz. A 37-point linear phase finite impulse response (FIR) filter has been used. This filter has normalised transition width $\Delta f = 0.1$, and 60 dB attenuation in the stop band [131]. The pitch period computation is performed at each 6.4 msec, using a block of 25.6 msec. The first stage of processing is the computation of the clipping level C_1 for the current 256 samples of speech. The clipping level is set at a value which is 50 percent of the smaller of the peak absolute sample values in the first (IPK1) and the last (IPK2) 8.5 msec portions of the block. Following that, the entire block (256 samples) is centre clipped, and then infinite peak clipped, resulting in a signal which assumes one of three possible values, +1 if the sample exceeds the positive clipping level, -1 if the sample falls below the negative clipping level, and 0 otherwise.

Following the clippings the autocorrelation function for the block is computed over a range of lags from 30 samples to 150 samples (i.e., from 3 msec to 15 msec period). The results are normalised by the autocorrelation at zero delay. The normalised autocorrelation function is then searched for its maximum value, and the position (IPOS) and value of the maximum value or the pitch peak (PPK) are sent to the V-UV detector. In general, if PPK exceeds a certain value, the block is classified as voiced and IPOS is taken as the pitch period, otherwise it is classified as unvoiced.

5.2.3 Initial Voiced-Unvoiced Classification

With silence segments removed from consideration (Section 5.2.1), the remainder of the utterance is segmented into voiced and unvoiced portions. The energy, the normalised first shift autocorrelation coefficient, and the pitch are computed for each frame of 6.4 msec. The energy is given as:

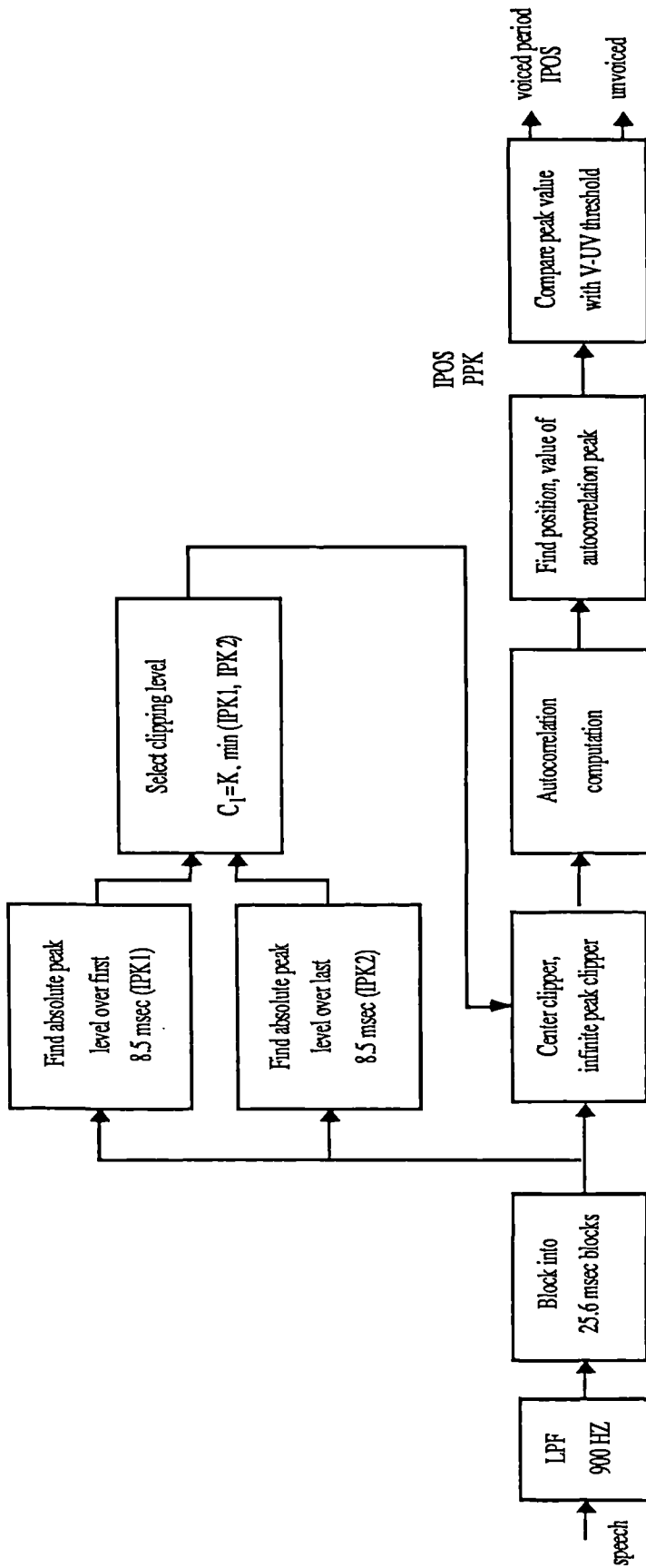


Fig. 5.3 Block diagram of the autocorrelation pitch detector.

$$E = \frac{1}{N} \sum_{n=1}^N s^2(n) \quad (5.2)$$

The normalised first shift autocorrelation coefficient is given as:

$$RN = \frac{\sum_{n=1}^{N-1} s(n) s(n+1)}{\sum_{n=1}^N s^2(n)} \quad (5.3)$$

The pitch detector (given in Section 5.2.2) provides the maximum normalised value PPK, and the position of the peak IPOS.

The decision algorithm is summarised in the chart of Figure 5.4 . The algorithm distinguishes between two main cases according to the energy value, E, of the frame. If E exceeds a certain threshold Vth (voiced threshold), the frame has a high chance of being tagged as voiced rather than unvoiced, but it has to satisfy certain further conditions. So, if the value of the normalised first autocorrelation coefficient RN is greater than or equal to 0.6, or PPK is greater than or equal to 0.45, then the frame is tagged voiced, otherwise it is tagged unvoiced. This is actually the case of vowels and semivowels. Now, if E is below Vth, the frame is more likely to be unvoiced, unless it passes one of the following tests:

- RN ≥ 0.9 and PPK ≥ 0.3
- RN ≥ 0.8 and PPK ≥ 0.45

The first condition is actually used for nasals and liquids, while the second is meant for voiced fricatives where fricative sounds might have high values for RN, but they can not have high peaks in the pitch detector. The Vth is taken equal to 10000. All the above thresholds are chosen empirically from observations of results from a large number of speech frames.

The V-UV-S decision is coded by the values: 1, 2, and 3 for voiced, unvoiced, and silence respectively.

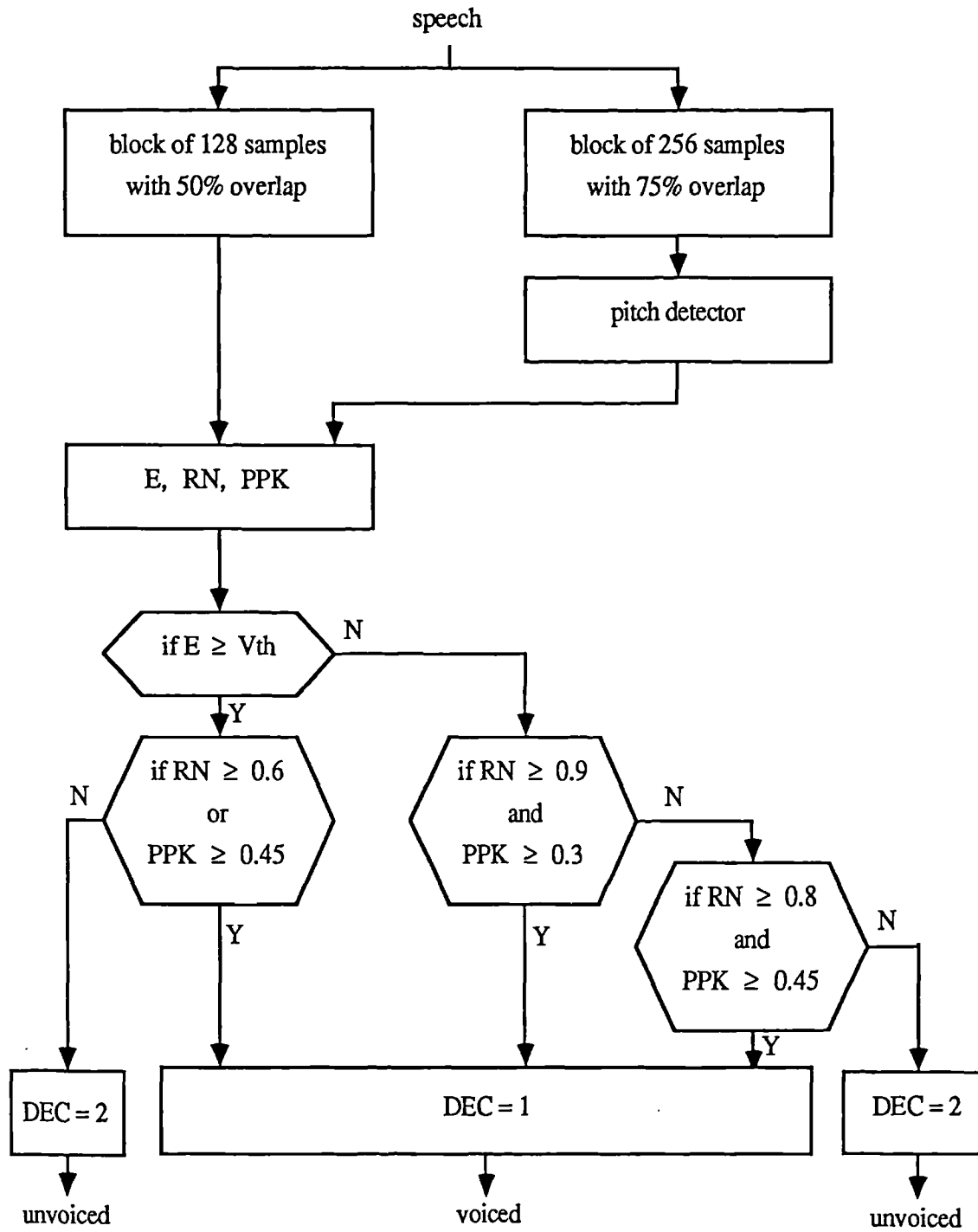


Fig. 5.4 Flow chart of the initial voiced-unvoiced decision

5.2.4 Weak Fricative Detection

As has been demonstrated in the initial silence detection (Section 5.2.1), a simple energy measure is used. This measure is not sufficient for separating weak fricatives (such as: /f/, /h/, /θ/, and /h/), from the background noise especially at word boundaries (i.e., the first or the last phoneme in a word). A simple procedure is proposed to deal with this problem [43].

The implemented algorithm is described as follows. The energy E_h , in the range 300-4800 Hz (see Section 5.2.1), and the zero-crossing rate ZCR are used to detect the frames related to weak fricative phonemes. Thus a silence frame is checked, if its energy E_h exceeds a certain noise threshold N_{th} , and at the same time if its ZCR exceeds a certain threshold Z_{th} , the frame is relabelled as unvoiced frame. This test is applied to all silence frames along a word, and to the first 30 frames before the beginning and after the end of a word. Thus, the endpoints of a word are readjusted again according to the presence of weak fricatives.

The N_{th} and Z_{th} are computed as follows. Statistical measurements are taken during 50 frames of silence before speech to estimate the average energy of the background noise (E_{bn}), the mean zero-crossing rate (ZC_m) and the standard deviation of the zero crossing rate (σ_{zc}) during silence. The zero-crossing threshold and the noise threshold are defined as follows:

$$Z_{th} = \min \{ 20, ZC_m + 2 \sigma_{zc} \} \quad (5.4)$$

$$N_{th} = 2 \cdot E_{bn} \quad (5.5)$$

The Z_{th} has taken the value 20 all the time in our experiments, and N_{th} has taken the value 200.

5.2.5 Editing V-UV-S Decision

Considerable editing is done on the preliminary V-UV-S decision to eliminate any spurious decision which may occur within a specific segment. This is achieved by using a nonlinear smoothing method. This method uses what is called the median smoother [132].

In a 3-point median smoother, the output $y(n)$ is taken as the 3-point median of $x(n-1)$, $x(n)$, and $x(n+1)$, i.e., the middle value when these three inputs are ordered in value.

As mentioned earlier, voiced, unvoiced, and silence frames are labelled or coded by the value '1', '2', and '3' respectively. A 3-point median can be, for example, applied to remove any voiced or unvoiced frame surrounded by silence frames, to remove any voiced or silence frame surrounded by unvoiced frames, or to remove any unvoiced or silence frame surrounded by voiced frames. Table 5.1 illustrates all possible cases encountered by the 3-point median. The last two lines (separated from other lines) in this table shows that the median smoother has failed to remove the unvoiced frame at time n surrounded by two voiced and silence segments (where a segment represents a string of similar frames). These two cases are tackled later on.

| inputs | | | | | outputs | | | | |
|--------|-----|---|-----|-----|---------|-----|---|-----|-----|
| n-2 | n-1 | n | n+1 | n+2 | n-2 | n-1 | n | n+1 | n+2 |
| 3 | 3 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 3 | 2 | 2 | 1 | 1 | 2 | 2 | 2 |
| 2 | 2 | 3 | 1 | 1 | 2 | 2 | 2 | 1 | 1 |
| 2 | 2 | 1 | 3 | 3 | 2 | 2 | 2 | 3 | 3 |
| 3 | 3 | 1 | 2 | 2 | 3 | 3 | 2 | 2 | 2 |
| 1 | 1 | 2 | 3 | 3 | 1 | 1 | 2 | 3 | 3 |
| 3 | 3 | 2 | 1 | 1 | 3 | 3 | 2 | 1 | 1 |

Table 5.1 Inputs and outputs of 3-point median

The V-UV-S decision along a word gives what is called the V-UV-S contour, which shows three levels '1', '2', and '3'. The V-UV-S contour is smoothed by running a 3-point, 5-point, and 7-point median. The 3-point median removes spurious segments of one frame duration, 5-point median remove spurious segments of two frames duration,

while the 7-point median remove spurious segments of three frames duration.

In all the above smoothers, the cases of having an unvoiced segment of one, two, or three frames duration surrounded by voiced and silence segments are not smoothed as is seen in the last two cases of Table 5.1. Actually the following case:

... 3 3 3 3 2 2 1 1 1 1 ... (silence - unvoiced - voiced)

represents the state of having a plosive phoneme followed by a voiced phoneme, where the plosive phoneme is represented by a silence gap followed by a short unvoiced segment representing the burst noise associated with such a phoneme. Fortunately, this case is not smoothed by the median smoothers. The other case:

... 1 1 1 1 2 2 3 3 3 3 ... (voiced - unvoiced - silence)

occurs within the transition period between either a voiced phoneme and a plosive one, or at the end of the word caused by the breathing noise.

Post-editing or post-processing is carried out in order to tackle some spurious segments which have not been removed by the median smoothers, and to handle those spurious segments having a duration of more than 3 frames. The following cases are dealt with by the post-editing process:

- An unvoiced segment of less than 5 frames' duration is relabelled as silence, if it is preceded by a voiced segment and followed by a silence segment.
- An unvoiced segment of less than 5 frames' duration is relabelled as voiced, if it is surrounded by voiced segments.
- A voiced segment of less than 5 frames' duration is relabelled as unvoiced, if it is surrounded by unvoiced segments.
- A silence segment of less than 7 frames' duration is relabelled as unvoiced, if it is not surrounded by voiced segments (i.e., between two unvoiced segments, or between voiced and unvoiced segments).

Figures 5.5 to 5.7 display a) the speech signal and b) the smoothed V-UV-S contour for three different words. The horizontal axes of both (a) and (b) graphs in each figure show the time along the word given in frame numbers (each frame equals to 6.4 msec). The

vertical axis of graph (a) shows the level of the speech samples which varies from -2048 to 2048 (12-bit), while the vertical axis of graph (b) shows three levels, i.e., 1, 2, and 3 relating to voiced, unvoiced, and silence decision respectively.

Figure 5.5b shows the V-UV-S contour for the word 'fataħa', where the word starts at frame number 27 and ends at frame number 138. This word has 7 segments which are in the following ranges (first frame-last frame):

- 27 - 47 unvoiced segment related to the phoneme /f/.
- 48 - 65 voiced segment related to the phoneme /a/.
- 66 - 77 silence segment related to the silence gap associated with the plosive phoneme /t/.
- 78 - 80 unvoiced segment related to the burst associated with the plosive phoneme /t/.
- 81 - 102 voiced segment related to the phoneme /a/.
- 103 - 120 unvoiced segment related to the phoneme /ħ/.
- 121 - 138 voiced segment related to the phoneme /a/.

The V-UV-S contour of this particular word gives most of the necessary information for the segmentation process, but unfortunately this is not always the case for most words.

Figure 5.6b shows the V-UV-S contour for the word 'kataba'. This word has 7 segments given as follows:

- 18 - 23 unvoiced segment related to the burst associated with the plosive phoneme /k/.
- 24 - 41 voiced segment related to the phoneme /a/.
- 42 - 49 silence segment related to the silence gap associated with the plosive phoneme /t/.
- 50 - 54 unvoiced segment related to the burst associated with the plosive phoneme /t/.
- 55 - 76 voiced segment related to the phoneme /a/
- 77 - 81 silence segment related to the silence gap associated with the plosive phoneme /b/.
- 82 - 109 voiced segment related to the phoneme /a/.

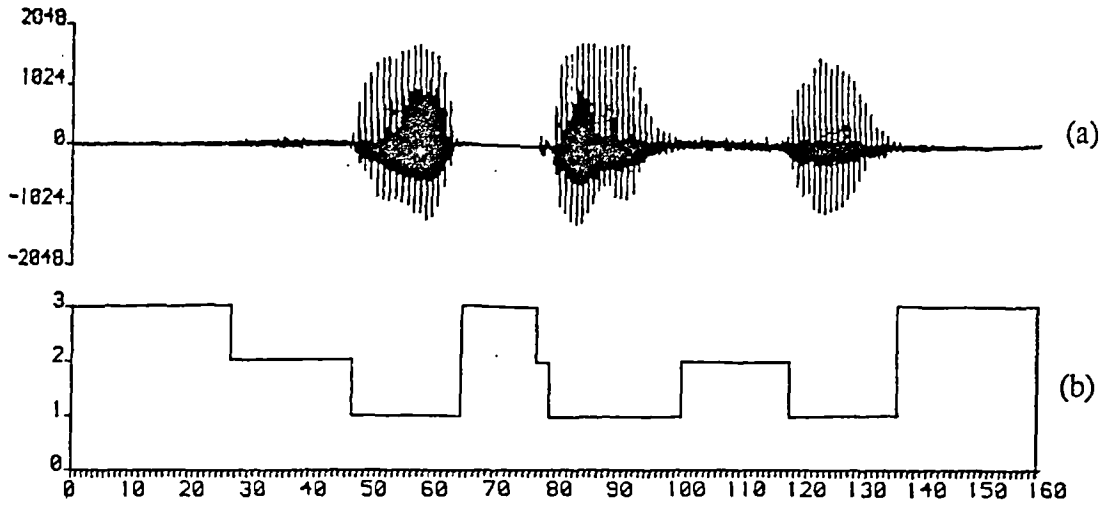


Fig. 5.5 The word 'fataħa', a) the speech signal, b) the V-UV-S contour

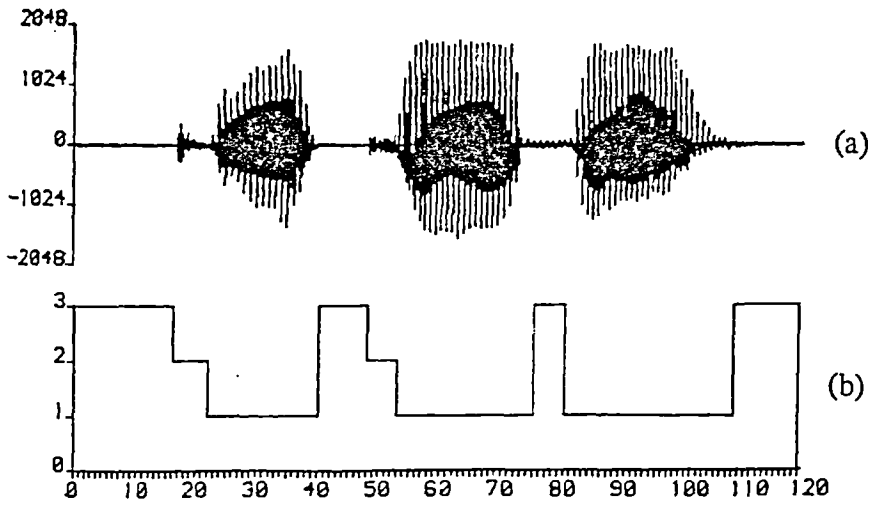


Fig. 5.6 The word 'kataba', a) the speech signal, b) the V-UV-S contour

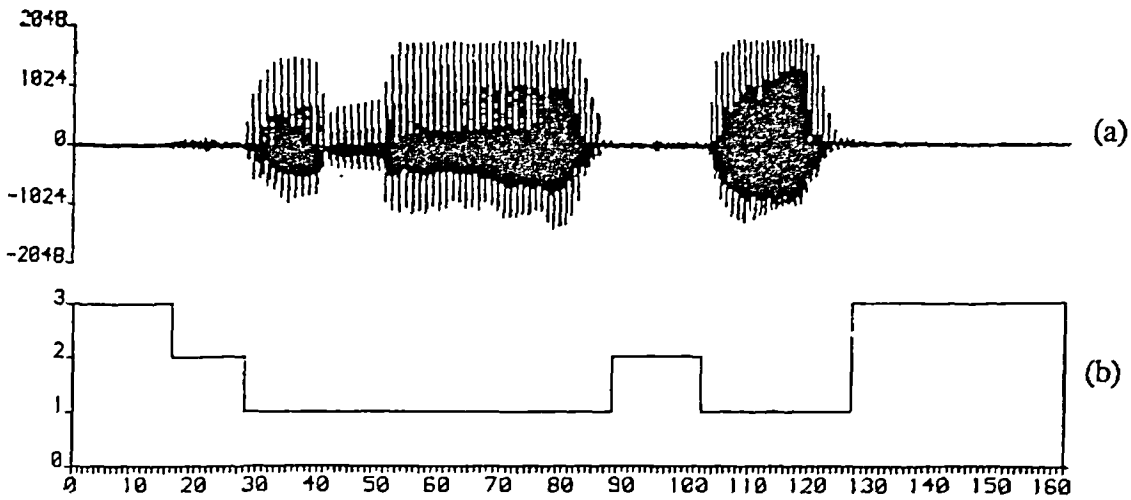


Fig. 5.7 The word 'ħalaaħa', a) the speech signal, b) the V-UV-S contour

Note that the V-UV-S contour indicates the level '3' or silence (silence segment between frames 77-81) for the voiced plosive /b/. The unvoiced segment related to the burst which is associated with plosive phonemes did not occur in this case, because this phoneme is surrounded by two vowels.

Figure 5.7b shows the V-UV-S contour for the word 'θalaaθa'. This word contains two weak fricative phonemes and four voiced phonemes. This word has four segments given as follows:

- 18 - 29 unvoiced segment related to the phoneme /θ/.
- 30 - 89 voiced segment related to the phonemes /a/, /l/, /aa/.
- 90 - 104 unvoiced segment related to the phoneme /θ/.
- 105 - 129 voiced segment related to the phoneme /a/.

Other V-UV-S contours for different words can be seen later on in the vowel recognition section and in Chapters 6 and 7.

5.3 Vowel Detection

It is well known that in most cases, vowels have more power than their adjacent consonants. The presence of vowels can be determined by looking for a local maximum of the log-energy contour, or of the loudness contour of the carrier word [41, 133].

The aim of this detection procedure is to determine the vowel steady-state regions, and to select reliable representative frames which are passed to the vowel identification procedure. The detection algorithm also has to define an estimated duration for each vowel, in order to distinguish between short and long vowels. This duration is also used in the segmentation procedure in Chapter 7.

5.3.1 Energy Peak Detection

The short-time energy (i.e., for a block of N samples) of the speech signal provides a convenient representation that reflects the amplitude variations of the speech signal. For the purpose of vowel detection, the energy is defined as:

$$ES = \left(\frac{1}{N} \sum_{n=1}^N S^2(n) \right)^{\frac{1}{2}} \quad (5.6)$$

where ES represents the square root of the energy over a block of N samples, and for simplicity it will henceforth be called energy. ES is computed every 6.4 msec, using a block of 128 samples (i.e., 12.8 msec) with 50% overlap over time for each word. The variation of ES over time gives the ES contour for each word. ES (square root of the energy) representation has actually given better results for vowel detection than were obtained using energy representation (on a linear or logarithmic scale).

The ES contour of each word is heavily smoothed (7 passes) in the time domain via a 3-point Hanning window (linear smoothing). The impulse response of this window (or filter) is:

$$\begin{aligned} h(n) = 0.5 & \quad \text{for } n = 0 \\ & 0.25 \quad \text{for } |n| = 1 \\ & 0 \quad \text{for } |n| > 1 \end{aligned} \quad (5.7)$$

The local peaks and valleys of an ES contour are then determined via a simple peak-picking algorithm. It has been noticed that most of the ES contour's peaks correspond to the vowel central regions. Post-editing is carried out to discard false and/or spurious peaks, and to detect clearly those prominent peak points which represent the vowel steady-state regions.

a) Initial Peak Detection

After some preliminary experimentation, it has been found that peaks are to be neglected if they satisfy any of the following conditions (see Figure 5.8):

- Peak within unvoiced or silence regions (segments)
- If $y(P) < ESt_h$
- If $[y(P1) - y(V1)] < (ESt_h / 2)$
- If $[y(P1) - y(V2)] < (ESt_h / 4)$

- If $y(P1) < (2 \text{ ESth})$ and $[y(P1)-y(V2)] < \text{ESth}$
- If $\{ [y(P1)-y(V1)] / [x(P1)-x(V1)] \} < (\text{ESth} / 20)$

where the threshold ESth is equal to 10% of the maximum value (ESmax) over the ES contour of each word.

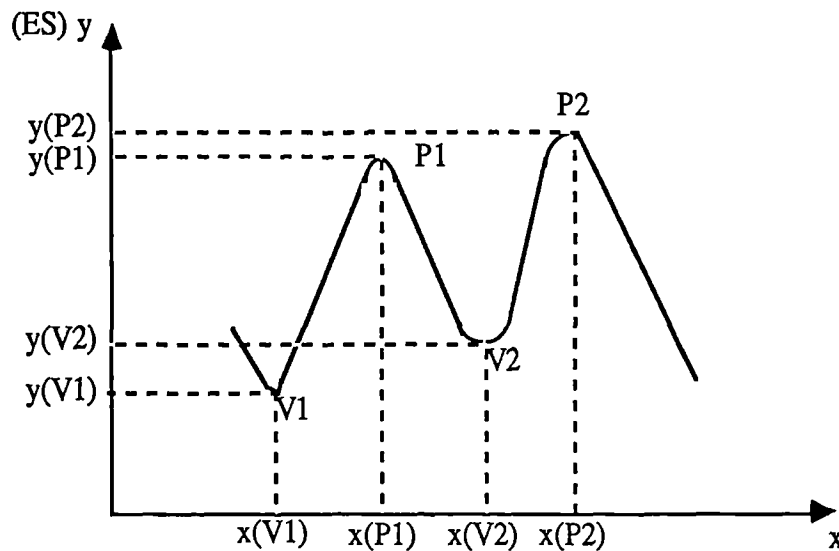


Fig. 5.8 Example of peaks and valleys along the energy contour (ES)

The first condition implies that only peaks in voiced regions are counted as candidates for vowel detection. The second condition leads to neglecting any peak whose value is less than ESth . When the difference between the energy of a certain peak and the energy of the preceding valley is less than half ESth , that peak is neglected according to the third condition. When the difference between the energy of a certain peak and the energy of the following valley is less than quarter ESth , that peak is neglected according to the fourth condition. The fifth condition leads to discarding any peak which has an absolute energy less than twice ESth and at the same time the difference between its energy and the energy of the following valley is less than ESth . The last condition is used to eliminate those peaks which might occur within the trailing consonants in the syllabic pattern /CVC/ (at the syllable boundary). Figure 5.9 Displays graphs for the word 'masaña', where a) shows the speech signal, b) the V-UV-S contour, c) the ES contour, and d) the loudness contour. The loudness contour is explained later on.

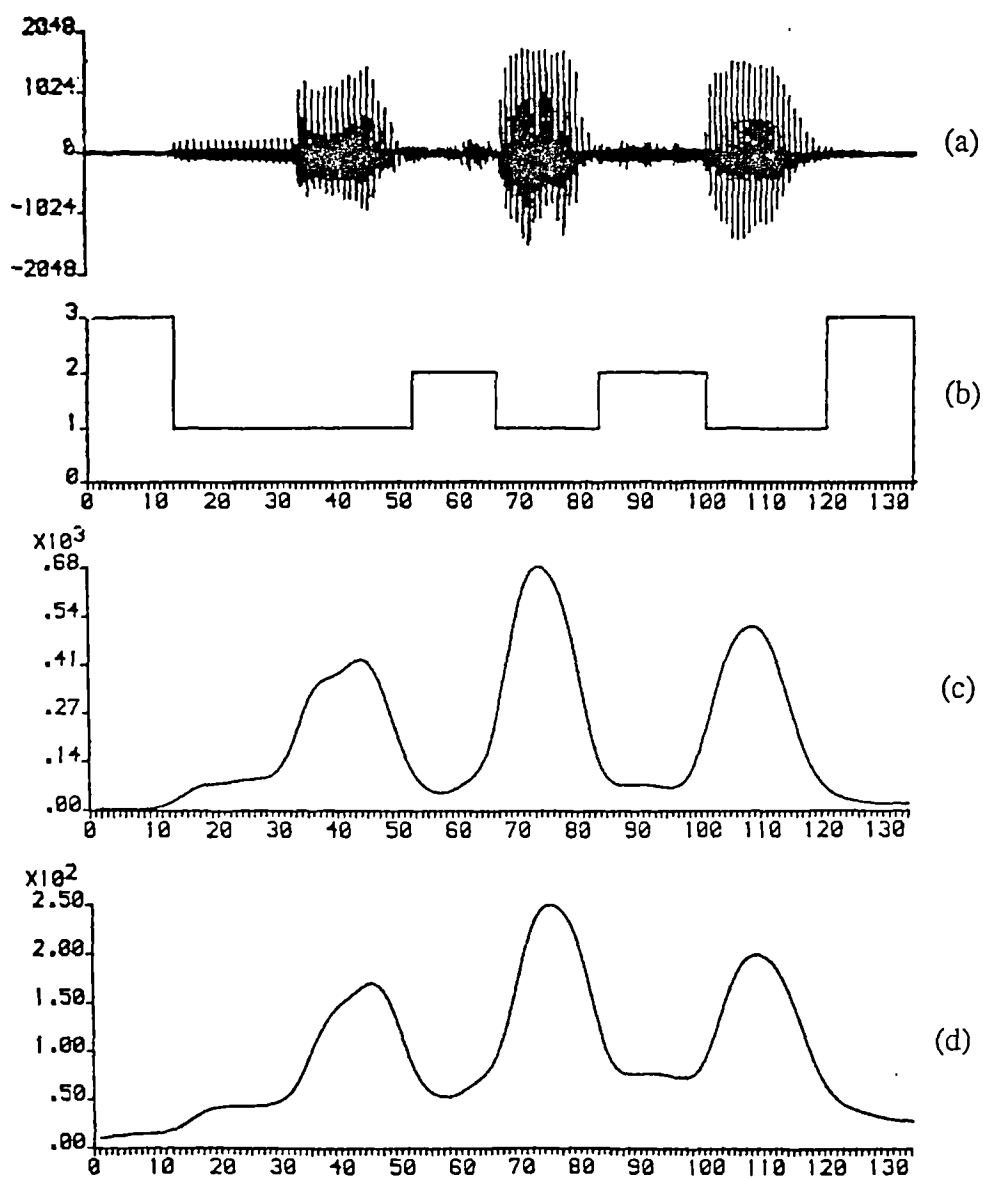


Fig. 5.9 Graphs for the word 'masaha'
 a) the speech signal, b) the V-UV-S contour
 c) the ES contour, d) the LO contour

b) Vowel Estimated Duration

The vowel estimated duration VED is defined as the duration of a vowel peak at 0.7 of its amplitude (i.e., at half the power). When the VED for a certain peak exceeds a certain threshold (VED_{th}), this peak is associated with a long vowel; otherwise it is associated with a short vowel. VED_{th} is taken as 22 frames or 140.8 msec. This threshold has been obtained statistically from an analysis results of the speech database, where VED took the values:

- 6 - 20 frames for short vowels (38.4-128 msec)
- 22 - 60 frames for long vowels (140.8-384 msec)

Actually, VEDs are related to the speaking rate. However, the VED_{th} could be estimated on-line from the relative duration of other short vowels in the same word. In this case, If the VED of a given vowel peak is more than 25 frames, the vowel is considered as a long vowel. But if it is less than or equal to 25 frames, we look at the VED of the preceding vowel (or at the VED of the following vowel if a preceding vowel does not exist), and the threshold is computed as:

$$\text{VED}_{th} = 1.6 \text{ VED} \quad (5.8)$$

For example, when the VED of the preceding vowel is 15 frames, the threshold VED_{th} is equal to 24 frames, but when it is 13 frames, the threshold VED_{th} is equal to 20 frames.

As was shown in Chapter 3, the duration of the Arabic vowel is very important, since increasing the duration of a vowel (while speaking) may lead (in most cases) to a word which has completely different meaning.

c) Eliminating False Peaks

Further processing is performed to remove false peaks in the ES contour. Figure 5.10 shows graphs for the word 'jafæaluuna', where a) displays the speech signal, b) the V-UV-S contour, and c) the ES contour. The second peak of the ES contour at frame 68 is a false peak, since it occurs within the duration of the voiced consonant /ε/. The fourth and fifth peaks are related to the long vowel 'uu'; such cases are discussed in the next section.

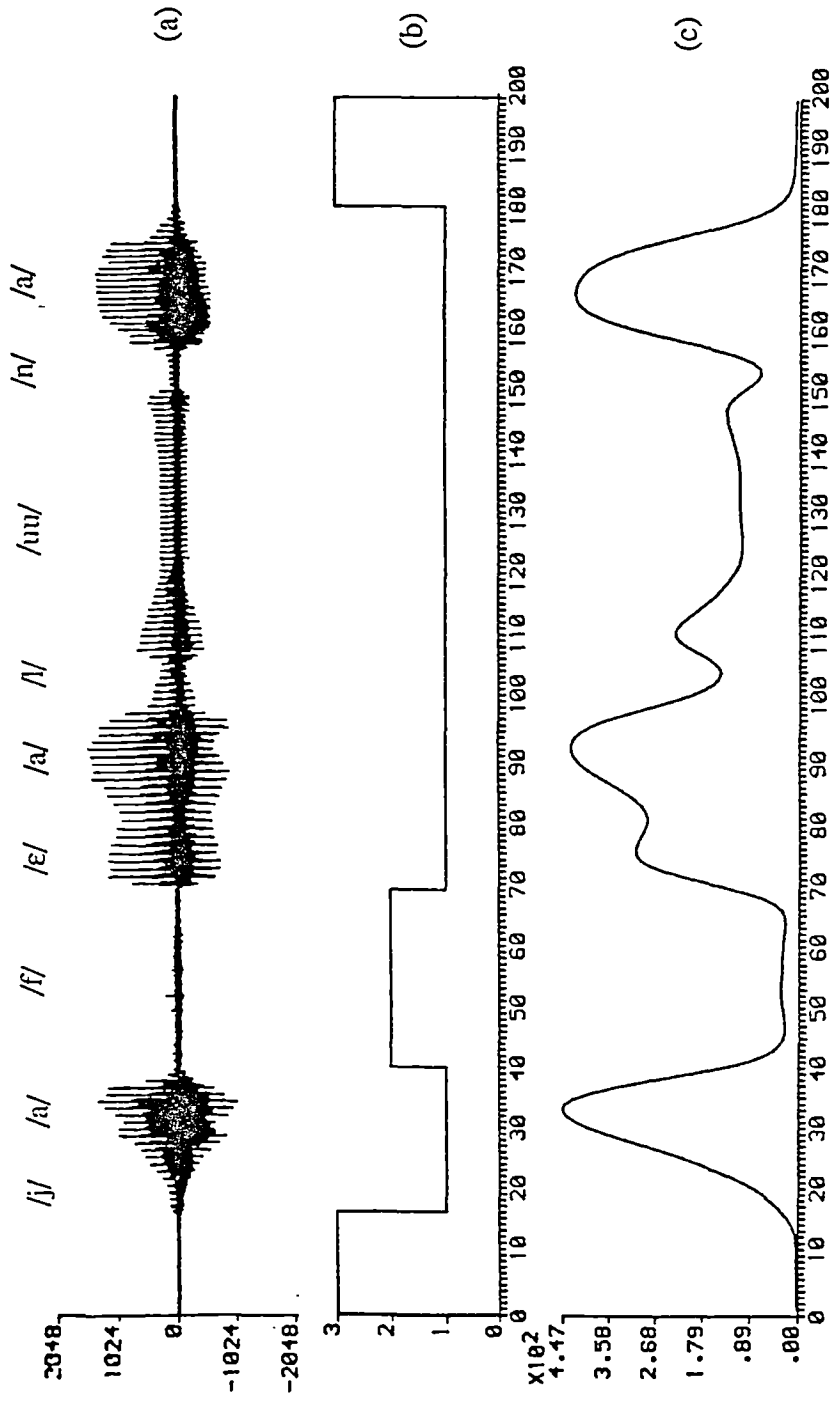


Fig. 5.10 Graphs for the word 'jafealuuna'
 a) the speech signal, b) the V-UV-S contour, c) the ES contour

Such a false peak is eliminated by passing the ES contour through the following test (see Figure 5.8):

$$[x(P2) - x(P1)] < 20 \text{ frames} \quad \text{and} \quad y(V2) > 0.75 \cdot \min \{y(P1), y(P2)\}$$

This condition states that, if the duration between two successive peaks is less than 20 frames, and at the same time the energy of the valley between them is more than three quarter of the energy of the lower peak, then the peak which has the lower value is neglected. Such false peaks have occurred in many cases, and for both consonants in the syllabic type /CVC/.

The first part of the above condition could be satisfied by two genuine successive peaks (i.e., where the duration between the centre of two vowels, or V-C-V, is less than 20 frames which could be called the minimum syllabic duration). But the second part of the condition is included to rule out such cases. For two genuine successive peaks, it was found that the amplitude of the valley between the two peaks (i.e., the consonant amplitude) always below 75% of the lower amplitude of the adjacent peaks (i.e., the vowel amplitudes), when the duration between the two peaks is less than 20 frames.

d) Long Vowel Detection

Figure 5.11 shows graphs for the word 'nuuhiĥaa' (the only word in Arabic which contains the three long vowels together), where a) displays the speech signal, b) the V-UV-S contour, and c) the ES contour. Here, the ES contour contains two peaks for each long vowel. Such a phenomenon is frequently associated with long vowels (see Figure 5.10c, where the fourth and the fifth peaks along the ES contour belong to the long vowel /uu/). Therefore, after detecting the prominent peaks, the possibility of having two adjacent peaks related to one long vowel is checked. In this respect, two cases are distinguished (see Figure 5.8):

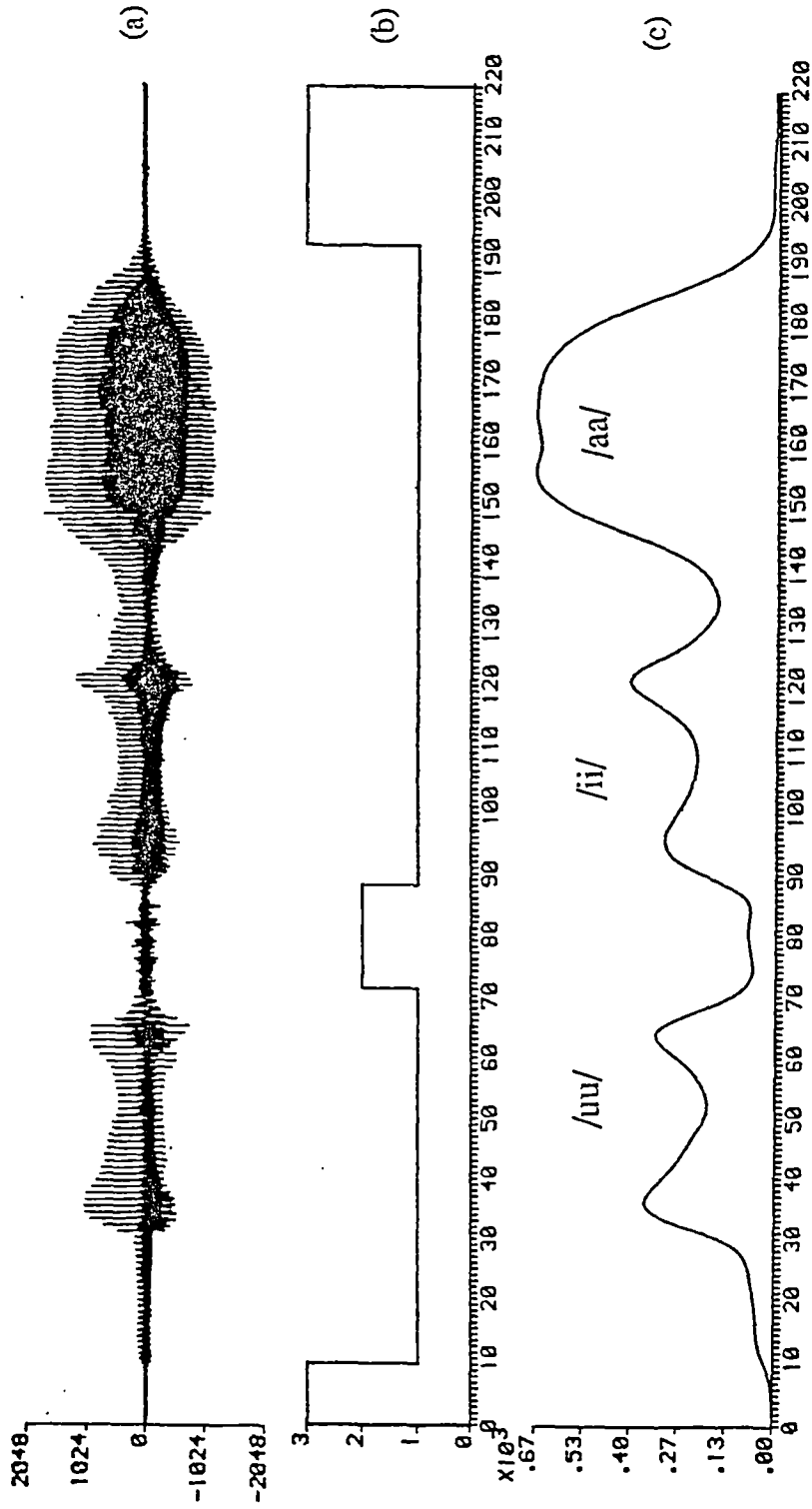


Fig. 5.11 Graphs for the word 'nuuhihaa'
a) the speech signal, b) the V-UV-S contour, c) the ES contour

- 1) If two successive peaks P1, P2 and the valley between them V2 satisfy the following condition:

$$\begin{aligned} \text{VEDs of both P1 and P2} &< 22 \text{ frames} \\ \text{and } [x(P2) - x(P1)] &> 20 \text{ frames} \\ \text{and } y(V2) &> 0.75 \max \{y(P1), y(P2)\} \end{aligned}$$

then both peaks belong to one long vowel, and are replaced by a new peak situated at the centre of the duration between the two peaks as follows:

$$x(P) = [x(P2) - x(P1)] / 2$$

This case occurs mainly for the long vowel /aa/.

- 2) If two successive peaks P1, P2 and the valley between them V2 satisfy the following condition:

$$\begin{aligned} \text{VEDs of both P1 and P2} &< 22 \text{ frames} \\ \text{and } [x(P2) - x(P1)] &\in [21 - 60] \text{ frames} \\ \text{and both } y(P1) \text{ \& } y(P2) &< \alpha \text{ ESmax} \\ \text{and } y(V2) &> 0.5 \min \{y(P1), y(P2)\} \\ \text{and } |y(P1) - y(P2)| &< 0.6 \max \{y(P1), y(P2)\} \end{aligned}$$

then both peaks belong to one long vowel, and are replaced by a new peak situated at the centre of the duration between the two peaks as follows:

$$x(P) = [x(P2) - x(P1)] / 2$$

This case occurs mainly for the long vowels /uu/ and /ii/.

α is taken as 0.75 if ESmax is more than a certain threshold (400), otherwise $\alpha = 1$ (this is the case when the test word does not contain any of the vowels /a/ or /aa/).

The above-mentioned conditions were found heuristically, and the thresholds were estimated (statistically) from observations of many ES contours of several words. Finally, it can be seen from Figure 5.11c that the variation in energy along the vowels /ii/ and /uu/ is higher than that along the vowel /aa/.

The VED of a long vowel which has two peaks is taken from the point (before the left peak) which has a value equal to 0.7 of the left peak value, to the point (after the right peak) which has a value equal to 0.7 of the right peak value.

e) Representative Frame Selection

Each vowel is represented by a single frame. This frame has to be present within the vowel's steady-state region where formant frequencies are almost at their nominal values for that vowel. The vowel representative frames (VRFs) are chosen initially at the location of the energy peaks along an ES contour.

Figure 5.12 display graphs for the word 'jamsaɦu', where a) shows the speech spectrogram of this word, b) the speech waveform, c) the V-UV-S contour, and d) the ES contour. The syllabic structure of this word is /CVC-CV-CV/. The first peak of the ES contour refers to the first vowel /a/ in the syllable /CVC/, the second peak refers to the second vowel /a/ in the syllable /CV/, and the third peak refers to the third vowel /u/ in the syllable /CV/. It can be seen that the ES contour displays three different shapes (bell shapes) for these three vowels. The shape related to the second vowel in the word is almost symmetric around the second peak, while the other two shapes are not symmetric. The slope of leading edge of the bell shape related to the third vowel is sharper than that of its trailing edge, and the vowel peak is situated at the beginning of the vowel, and almost within the transitional portion between the two successive phonemes /ɦu/ (see the spectrogram of this word in Figure 5.12a). In this case an error would occur during the vowel identification phase if the vowel representative frame is taken at this peak. Instead of that, the representative frame is chosen at the middle of the vowel estimated duration VED. The peak point related to the second vowel on the ES contour coincides with the central point of the vowel estimated duration (VED). From the bell shape of the first vowel, it can be seen that the slope of its leading edge is lower than that of the trailing edge. The spectrogram of this word shows that if the VRF is taken at the middle of the VED of the first vowel, an error would occur. The first vowel /a/ is preceded by a semivowel /j/, and the second formant F2 moves from above 2000 Hz within the duration of the phoneme /j/ towards the nominal value (1500 Hz) of the vowel /a/, and it almost reaches this value at the end of the vowel duration where it coincides with the peak location of this vowel on the ES contour. For this reason, the VRF of such a shape is taken exactly at the peak of the ES contour which is related to this vowel. For the second and the third shapes (related to the second and the third vowels), the VRFs are taken at the middle of the VEDs.

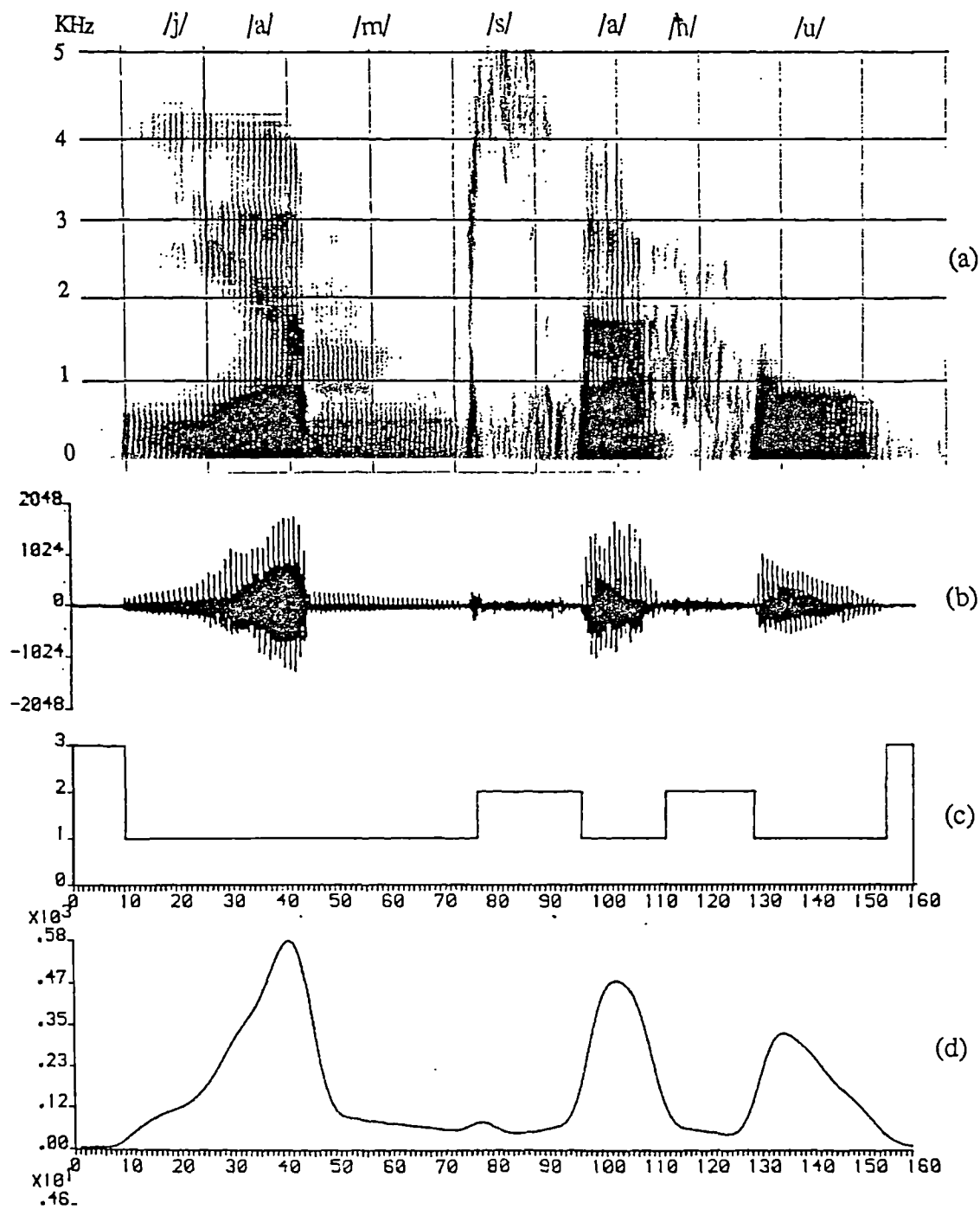


Fig. 5.12 Graphs for the word 'jamsaḥu'
 a) the speech spectrogram, b) the speech signal
 c) the V-UV-S contour, d) the ES contour

As can be seen, the ES contour of the word 'jamsaḥu' displays three different shapes for the three vowels in this word. The first shape usually occurs when a vowel is preceded by a semivowel. The third shape occurs in the realisation of the vowels /i/ and /u/, which usually have low energy compared with the vowel /a/, and whose peaks on the ES contour are towards the preceding consonant in their carrier syllables.

The VED is recalculated at 0.7 of the amplitude at the location of the VRF for both short and long vowels having only one peak.

5.3.2 Loudness Peak Detection

Loudness is defined as that attribute of auditory sensation in terms of which sound may be ordered on a scale extending from soft to loud. The loudness contour along a word was used to locate the syllable nucleus or the vowel [41]. The total loudness is obtained as the summation of specific loudness, extending over a bank of filters. The specific loudness is computed as follows.

The speech signal is passed through a bank of K filters which are linearly spaced below 1 KHz and logarithmically spaced above 1 KHz (up to 4.8 KHz). The computation of the filter bank parameters and of the energy in each channel is explained in detail in Chapter 6. The output energies of the K channels are weighted by an equal-loudness curve which approximates the auditory response over mid-range intensity levels (the weighting curve is actually the inverse of the hearing sensitivity curve at mid-range intensity level). This equal-loudness curve has a slope of +10 dB/oct in the range 0.1-0.4 KHz, flat in the range 0.4-1.2 KHz, +6 dB/oct in the range 1.2-3.1 KHz, and flat in the range 3.1-5 KHz. Then the cube root (Stevens' power law [134]) of the weighted energy for each channel is taken to obtain the specific loudness. The total loudness is given as the summation of the specific loudness along the K channels of the filter bank.

The total loudness is computed along each word to yield the loudness (LO) contour. Smoothing and editing are also carried out on the LO contour, as in the case of the ES contour. Figure 5.9d shows the LO contour for the word 'masaha'. This contour shows no significant differences compared to the ES contour in Figure 5.9c, as far as the vowels are concerned. In the LO contour, consonants have relatively higher amplitude than in the ES contour because of the cube root operation applied to the energy of each

channel in the filter bank. Both contours have given almost the same results (for several test words), where different thresholds were used in the peak-picking and editing processes of the two contours.

However, the ES contour is considered for vowel detection in this research work, since it requires simpler and faster computation.

5.2.3 Results of the Vowel Detection

The vowel detection algorithm introduced in the previous section has been quite successful. However, there were specific cases in which the detection algorithm failed. For example, Figure 5.13 displays graphs for the word 'jusaawii', where a) shows the speech signal, b) the V-UV-S contour, c) the ES contour, and d) shows what is called the spectral variation contour (SV). The SV contour displays peaks at the transition between successive sub-word units. The SV contour is explained in detail in Chapter 6. The ES contour of this word shows no prominent peak (or peaks) for the long vowel /ii/ at the end of the word. This vowel has energy less than the preceding consonant. The SV contour shows peaks at the transition between adjacent phonemes in this word, and the last two peaks refer to the boundaries of the vowel /ii/. By combining information from the V-UV-S contour, the ES contour, the SV contour, durational information, and phonological constraints (such as allowable syllabic structures), the missing vowel can be recovered as it is demonstrated in Chapter 7. For another example, the ES contour of the word 'ʔarbaæa' shows no peak related to the last vowel in this word (see Figure 7.18 in Chapter 7). This occurs because the voiced consonant 'æ' has a relatively higher energy than the following vowel /a/. This problem is also dealt with in the segmentation and error correction procedures given in Chapter 7.

Most of the errors occur mainly when a low energy vowel such as /i/ or /u/ appears after or before voiced consonants such as /n/, /m/, /w/, /ɛ/, etc., where the energies of these consonants are of the same order as the energy of the vowel.

Some of the errors were made by the speakers themselves, where an artificial pause was encountered between the consonant and the vowel of the syllable /CV/, leading to two distinct peaks in the ES contour. This case is also avoidable by asking the speaker to speak fluently (not artificially) with normal speed.

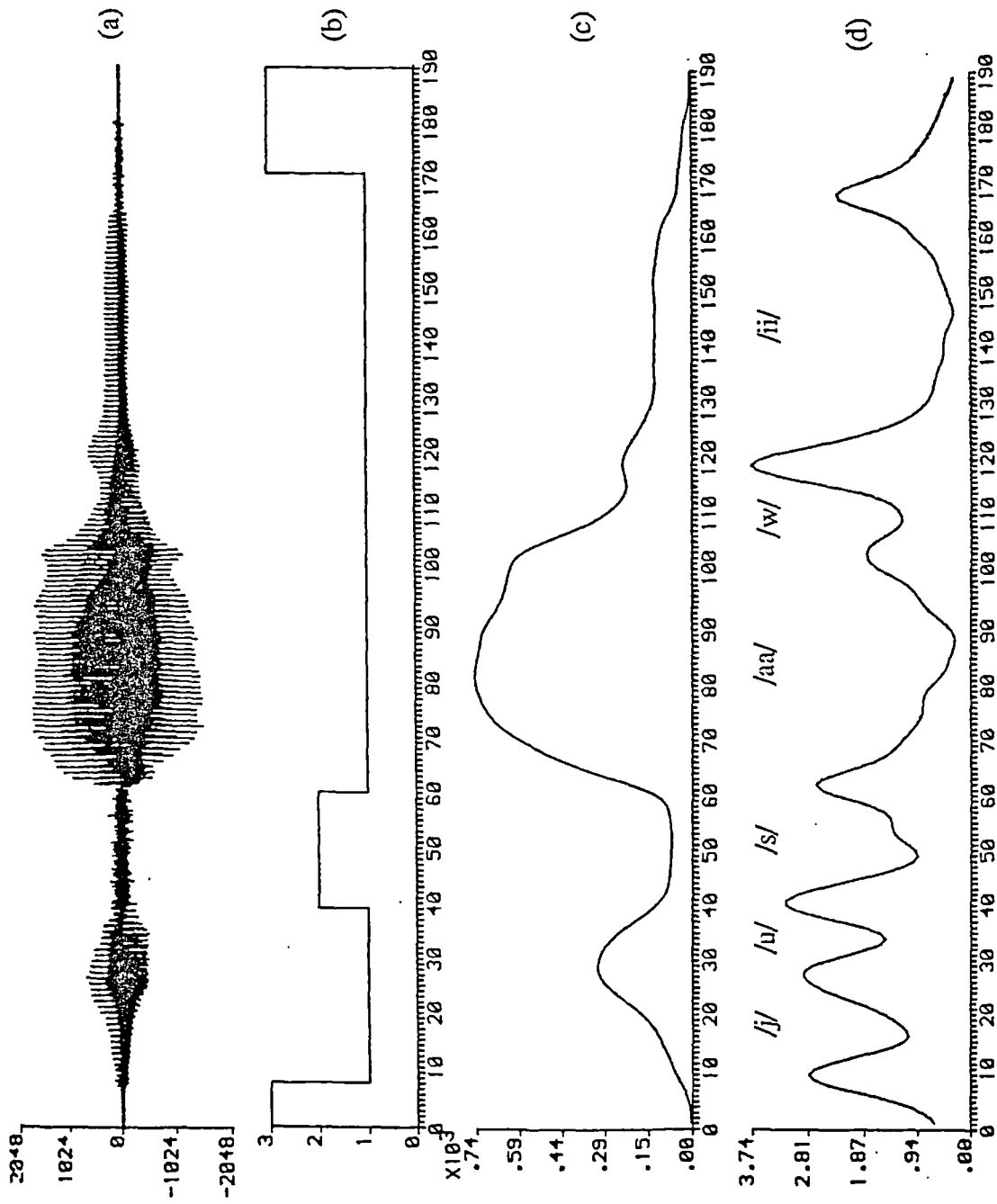


Fig. 5.13 Graphs for the word 'jusaawii'

a) the speech signal, b) the V-JV-S contour, c) the ES contour, d) the SV contour

A set of 100 different words [see appendix A] has been used to test the performance of the implemented algorithms in the proposed speech recognition model. The 100 words or a subset of them were uttered by five speakers are as follows:

- speaker (male) MZ 200 words (two repetition of the 100 words)
- speaker (male) YM 200 words (two repetition of the 100 words)
- speaker (male) HT 100 words (the set of 100 words)
- speaker (male) MB 50 words (50 words out of the set of 100 words)
- speaker (female) HK 20 words (20 words out of the set of 100 words)

Thus, the speech database used to test the vowel detection procedure comprises 570 words. Only 300 words uttered by three speakers (MZ, YM, HT, 100 words each) have been used for the extraction of the system parameters and thresholds.

The set of 100 words contains 272 vowels distributed as follows:

| | |
|-----------|------------|
| /a/ : 162 | /aa/ : 141 |
| /u/ : 22 | /uu/ : 7 |
| /i/ : 30 | /ii/ : 10 |

The 570 test words contains 1551 vowels distributed as follows:

| | |
|-----------|------------|
| /a/ : 927 | /aa/ : 231 |
| /u/ : 120 | /uu/ : 44 |
| /i/ : 170 | /ii/ : 59 |

The mispronunciation error rate was about 2%, and could be removed by proper pronunciation, while the unavoidable error rate (due to the structure of the test words, e.g., having a high energy voiced consonant followed by a low energy vowel at the word-final position) was about 1%. Some of these errors are tackled in the segmentation and error correction procedures given in Chapter 7, where a simple linguistic knowledge is employed.

Finally, the output of the detection stage is a representative frame of 256 samples for each detected vowel, which is passed to the identification stage. Also, this stage provides the vowel estimated duration VED. This VED is used later on to decide whether the vowel is short or long, and to aid in the vowel correction algorithm as will be explained in Chapter 7.

5.3 Vowel Identification

The aim of the vowel identification is to determine the vowel type using a single representative frame which is the output of the vowel detection stage. Also, the VED is used to decide whether the vowel is short or long. Vowel identification is achieved by two techniques, i.e., vector quantisation and formant methods.

5.3.1 Vector Quantisation Method (VQ)

In this method, the speech samples of the vowel representative frame are pre-emphasised using a first-order filter with transfer function:

$$H(z) = 1 - 0.95 Z^{-1} \quad (5.9)$$

passed through a Hamming window, and a 16th order LPC analysis is performed using the autocorrelation method [30]. Then, 18-LPC-derived cepstral parameters are computed as follows:

$$LPCC_i = LPC_i + \sum_{k=1}^{i-1} \frac{k-1}{i} LPCC_{i-k} \cdot LPC_k \quad (5.10)$$

where N is the number of cepstral parameters and $i = 1, 2, \dots, N$. K is the LPC model's order and $k = 1, 2, \dots, K$. Any LPC coefficient whose index is above the model's order, is taken as zero. As a result, each vowel is represented by a vector of N LPC-derived cepstral parameters.

The idea of this method is to design a codebook for the training vectors of each vowel type. An unknown vowel vector is handled by each vowel vector quantiser and the minimum VQ distance for each codebook is computed. The recognised vowel is chosen as the one whose VQ distance is minimum.

The design of a vector quantiser is summarised as follows. Assume that a training set $\{T\} = \{T_1, T_2, \dots, T_K\}$ of K cepstral vectors is given. It is desired to create a codebook of M vectors such that the average distance of a vector in $\{T\}$ from the closest codebook entry (codeword) is minimised. Thus, we wish to find a set $\{R\} = \{R_1, R_2, \dots, R_M\}$

of reference vectors that minimises the average distance given by:

$$D_K(M) = \min_{\{R\}} \left[\frac{1}{K} \sum_{i=1}^K \min_{1 \leq m \leq M} [d(T_i, R_m)] \right] \quad (5.11)$$

where $d(T_i, R_m)$ is the Euclidian distance between a training vector T_i and a codebook entry R_m . The optimum codebook is generated by minimising the distortion expressed in Eq. (5.11) over a large number of training vectors through an iterative process. This equation can be solved efficiently by the so called binary-split algorithm [135-137].

a) The Binary-Split Algorithm

This algorithm begins by finding an optimum solution for a codebook with two entries (i.e., $M = 2$), starting with an initial guess of two vectors (or using the centroid of the entire training set). The optimal solution (optimal codeword) is reached when the rate of decrease in the average distortion $D_K(M)$ of the K training vectors satisfies a predetermined threshold. Then, each optimal codeword is split into two ($M = 2 \cdot M$), and used as an initial guess for the design of a codebook of 4 entries. The binary split continues until the number of entries is equal to the desired codebook size.

The algorithm can be described in the following steps:

step 1:

- Compute the centroid C of the entire training sequence $\{T\}$.
- Split this centroid into two close vectors, i.e., $R_1 = C \cdot (1 - \alpha)$ and $R_2 = C \cdot (1 + \alpha)$, where α is a fixed perturbation vector.
- Set $M = 2$ and set the initial average distortion (D_{old}) of the training vectors to a large value.

step 2:

- Given M codewords, the training vectors $\{T\}$ are grouped into M clusters, where each training vector is assigned to the codeword closest to it by computing the Euclidian distance between this vector and T_i and the codeword R_m as follows:

$$d(T_i, R_m) = \sum_{n=1}^N (T_i(n) - R_m(n))^2 \quad (5.12)$$

step 3:

- Compute the average distortion $D_K(M)$ of the training K vectors which have been assigned to M clusters according to the following equation:

$$D_K(M) = \frac{1}{K} \sum_{i=1}^K \min_{1 \leq m \leq M} d(T_i, R_m) \quad (5.13)$$

step 4:

- If the percent change in the distortion (i.e., $\text{DIST} = |D_K(M) - D_{\text{old}}| / D_{\text{old}}$) is not less than a preset value ϵ then:
 - Set $D_{\text{old}} = D_K(M)$
 - Update the centroids of the M clusters (R_1, R_2, \dots, R_M).
 - Go to step 3

This process is iterated until DIST is less than ϵ .

step 5:

- If M is less than the desired size, split each centroid into two close vectors by multiplying each centroid with the values $(1 - \alpha)$ and $(1 + \alpha)$, where α is a fixed perturbation vector. M becomes double the previous M . Then, the entire process is repeated until M becomes equal to the desired codebook size.

The centroid vector of a given cluster is computed as the average of the cepstral coefficients of all the vectors in that cluster. The flow chart given in Figure 5.14 summarises the binary-split algorithm. The above described algorithm is actually called the full search binary-split VQ algorithm, where each vector in the training set is compared with every codebook entry. The result of this algorithm is a codebook of M entries (or codewords) which represents the centroids of the M clusters.

In our implementation, ϵ was chosen as 0.5 per cent (or 0.005), and α is taken as 0.01.

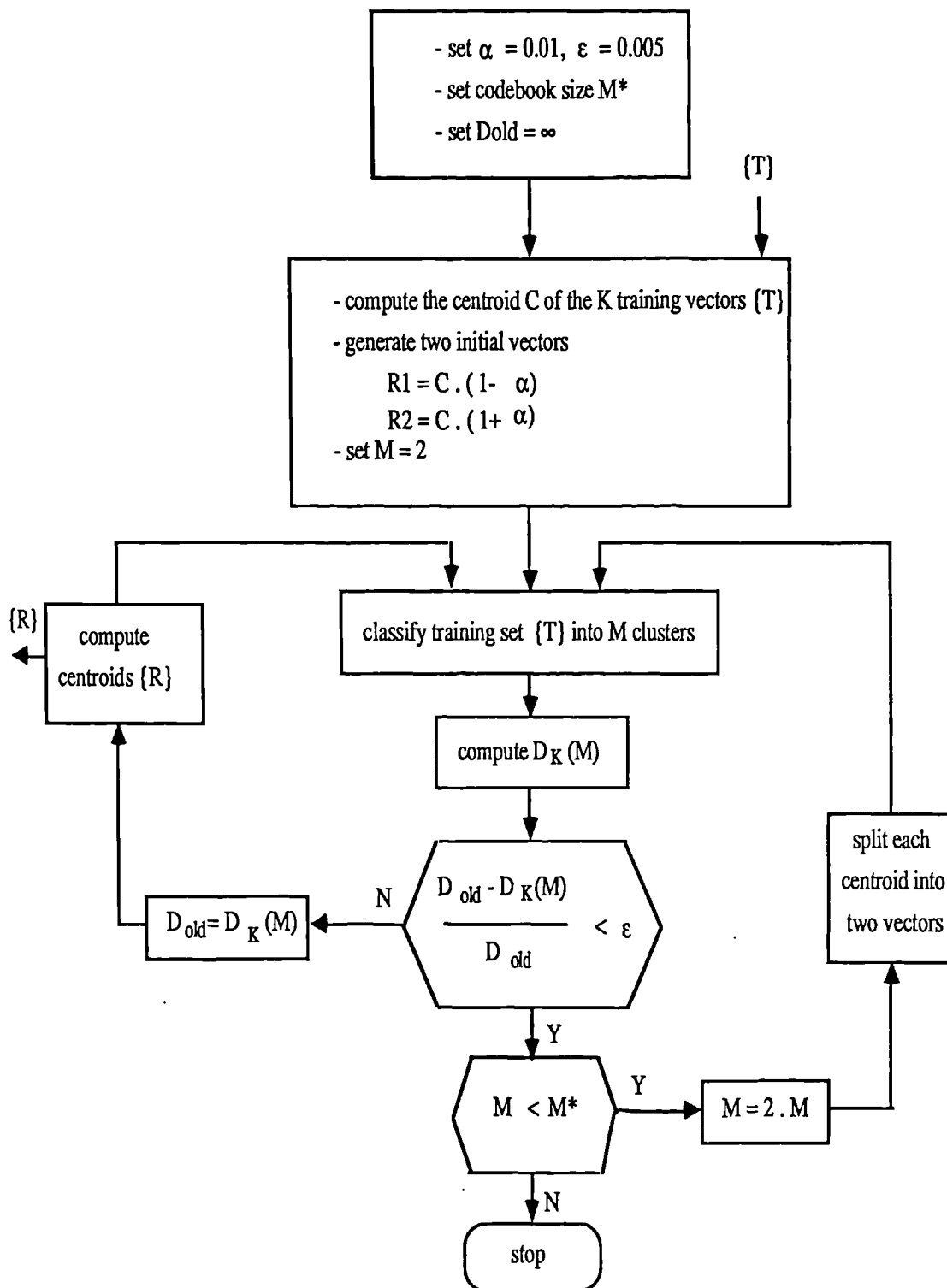


Fig. 5.14 Flow chart of the full search binary-split VQ algorithm

b) Experimental Results

As shown earlier, Arabic has six vowels divided into two sets, i.e., three short /a/, /u/, and /i/, and three long /aa/, /uu/, and /ii/. Ignoring the vowel length, we can say that we have three vowel types or groups, i.e. /a/, /u/ and /i/.

The training set used in the codebook design consists of 816 cepstral vectors, where each vector is derived from the vowel representative frame (during the vowel detection phase). 816 vowels (272 x 3) are extracted from the speech database of 100 words uttered by 3 male speakers. These vowels are distributed as follows:

| | | |
|------------|-----------|-----------|
| /a/ : 486 | /u/ : 66 | /i/ : 90 |
| /aa/ : 123 | /uu/ : 21 | /ii/ : 30 |
| ----- | ----- | ----- |
| 609 | 87 | 120 |

The following two tests have been performed:

- In the first test, the six vowels were represented by six different codebooks of size 16. In a series of recognition experiments the error rate was about 20 %. This error occurred mainly between the short and long vowels of the same type (e.g., vowels of the type /aa/ assigned to the type /a/ and visa versa).
- In the second test, the training data of short and long vowels of the same type (e.g., /aa/ and /a/), are grouped together to yield three different groups corresponding to the three vowel types. Then, codebooks of 8, 16, and 32 entries per vowel have been designed. The recognition results are summarised in Table 5.2.

Table 5.2 shows that the error rate is relatively small. The error rate went down from 4% to 1% when using a codebook of 32 entries for the vowels /a/ and /aa/. The remaining 1% of the errors occur mainly when the vowel /a/ is in association with the consonant /j/ in the syllable /ja/ (due to the coarticulation effect). The phoneme /j/ has a high second formant frequency (2400 Hz), and the vowel duration is not enough to allow F2 to go down to its nominal value for the vowel /a/ (about 1500 Hz) (see Figure 5.12a).

It is believed that in a multi-speaker system (used by a large number of male and female speakers), the error rate will increase because of the overlap between vowel parameters of the three vowel clusters across speakers. For this reason, vowel identification is performed by using another method which is based on the vowel formant frequencies.

| | V Q codebook size | | |
|-----------|-------------------|----|----|
| | 8 | 16 | 32 |
| /a/, /aa/ | 4% | 4% | 1% |
| /u/, /uu/ | 0 | 0 | 0 |
| /i/, /ii/ | 1% | 0 | 0 |

Table 5.2 Vowel identification error rate

5.3.2 Formant Method

In this method, vowels are represented by the first two formants F1 and F2. These formant frequencies are extracted from the vowel representative frame. Figure 5.15 displays a scatter plot for the six Arabic vowels in the F1-F2 plane (272 vowels by speaker MZ), where short vowels are represented by the symbol 'a', 'u', and 'i' and their long counterparts by the symbol /A/, /U/ and /I/. The figure shows an overlap between the short and the long vowel of the three vowel groups, whereas it is easy to discriminate between the three vowel groups (or clusters).

Figure 5.16 illustrates a flow chart of the vowel identification according to F1 and F2. Figure 5.17 shows the boundaries between the vowel areas in the F1-F2 plane (which are considered in the decision algorithm of Figure 5.16) for male speakers, (the numbers in brackets are for female speakers). This decision algorithm has given perfect accuracy for vowel identification in the speech of four male speakers (1496 vowels which are included in the 550 test words uttered by the four male speakers).

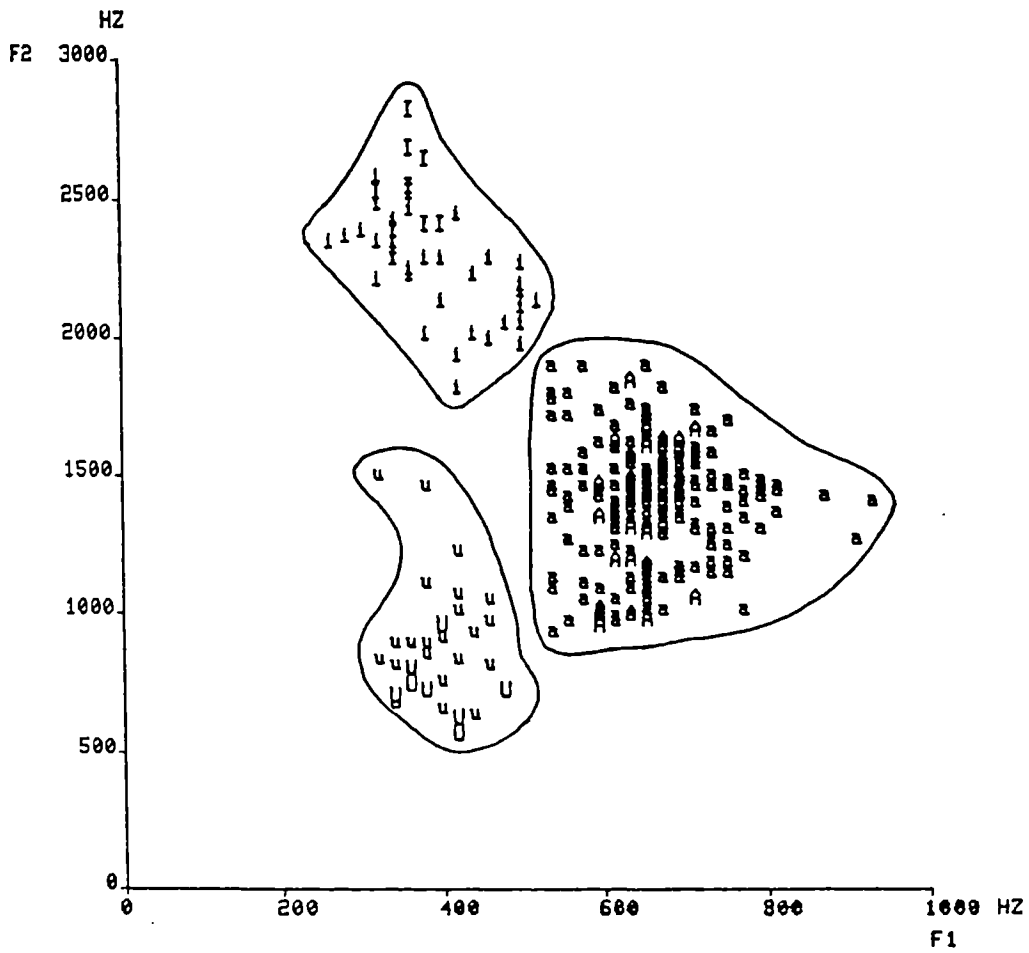


Fig. 5.16 Scatter plot of the six vowels in the F1-F2 plane
 single speaker, 272 vowels
 short vowels 'a', 'u', 'i'
 long vowels 'A', 'U', 'I'

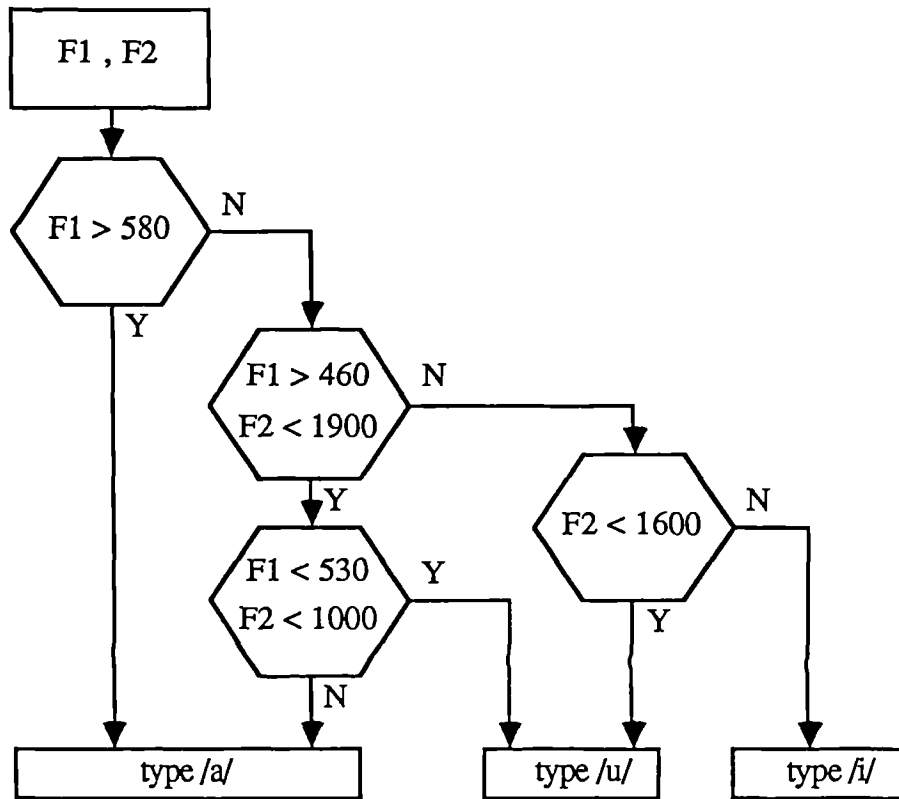


Fig. 5.16 Flow chart of the vowel identification using F1 and F2

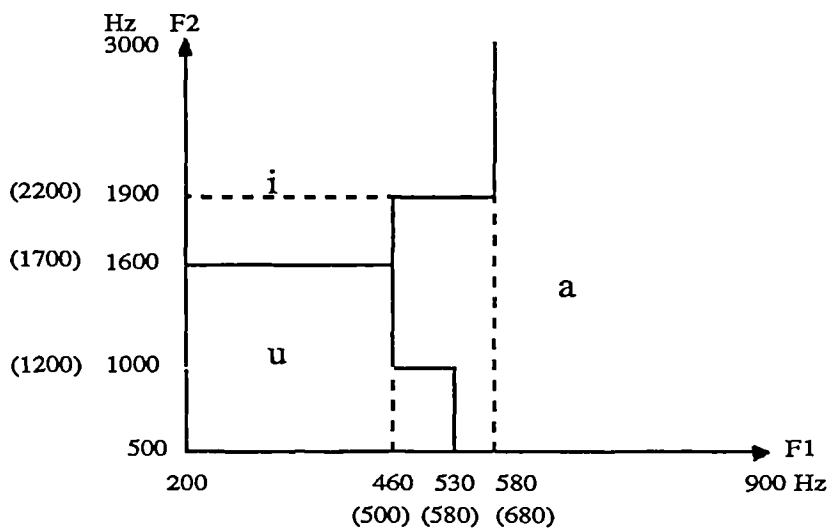


Fig. 5.17 The boundaries between the vowel areas in the F1-F2 plane
(the numbers in brackets are for female speaker, while the other numbers are for male speaker)

The vowel recognition algorithm (both the detection and identification phases) was tested using 55 vowels (in 20 words) uttered by a female speaker. The vowel detection phase gave correct results, while in the vowel identification phase different thresholds for F1 and F2 are used in the decision algorithm of Figure 5.16. These thresholds are taken from Figure 5.18 (the numbers in brackets).

a) Formant Normalisation

For multi-speaker systems (a large number of male and female speakers), the values of F1 and F2 may vary considerably across speakers, therefore formants should be normalised to account for these differences. One normalisation method is to normalise formants by the estimated vocal tract length of each vowel [138]. In this method, formant frequencies of the normalised vocal tract shape are computed by multiplying the normalised formant frequencies by the length factor L/L_R , where L_R is a reference length, and L is the estimated vocal tract length.

Vocal tract length computation is based on the idea that higher formant frequencies tend to be regularly spaced. By assuming that those higher formant frequencies do not deviate much from those of a uniform tube having the same length, the length L is estimated from the formant frequency F_i as :

$$L = \frac{(2i - 1) C}{4 F_i} \quad (5.14)$$

where C is the sound velocity (340 m/sec). F_i could be the fourth formant frequency or any other higher formant frequency. For example, if $F_4 = 3400$ Hz, L is equal to 17.5 cm. In this case L_R could be chosen as 17.5 cm.

It has been reported that this normalisation method has improved the accuracy of vowel identification process [138].

b) Formant Estimation

One way of estimating the formant frequencies of the speech signal over a short-time is through LPC analysis. The LPC analysis determines the coefficients (a_n 's) of the

autoregressive model (all-pole model) whose transfer function is given as:

$$H(z) = \frac{1}{1 + \sum_{n=1}^P a_n z^{-1}} \quad (5.15)$$

where P is the order of the prediction filter. A direct method for extracting formants is to determine the poles of the transfer function of Eq. (5.15). This involves solving the roots of a P th degree polynomial, which is computationally very tedious and requires high precision complex arithmetic. Another method often used is to locate the peaks of the LPC log-magnitude spectrum [30].

In our recognition system, a method based on the linear prediction phase spectrum has been implemented [139]. The log-magnitude spectrum of the LPC model ($\log(|H(f)|)$), shows peaks at resonant frequencies (formants). It has been shown [139] that a plot of the derivative of the phase spectrum (DPS) (i.e., the negative of the group delay function NGDF) of a resonance, closely resembles the shape of its magnitude spectrum. Hence, the frequency of a resonance can be estimated from the position of the peak of the NGDF, and the bandwidth of the resonance is proportional to the inverse of the height of the peak in the NGDF. An all-pole model can be regarded as a cascade of resonators. The overall phase spectrum of a cascade of resonances is a summation of individual phase spectra, hence each resonance curve will have very little influence on the shapes of other resonance curves. On the other hand, the overall magnitude spectrum is the product of individual magnitude spectra. This property makes the detection of closely spaced formants with different bandwidths easier from the NGDF (or DPS) rather than from the magnitude spectrum.

The NGDF is computed as follows:

- The speech signal of each vowel representative frame VRF (256 samples) is pre-emphasised by using a first order filter given in Eq. (5.9), is passed through a Hamming window, and then a P th-order LPC analysis is performed using the autocorrelation method [30].
- Compute 512-point DFT of the sequence $\{1, a_1, a_2, \dots, a_P\}$ appended with appropriate number of zeros using FFT algorithm.

- obtain the phase of the DFT components using the arctan function modulo- 2π .
 - Find the NGDF by computing the difference in phase between the successive points in the frequency domain. The absolute value of this difference is in the range $[-\pi, \pi]$.
- This NGDF will show peaks at the resonance frequencies (formants), due to the abrupt changes in the phase spectrum, since the arctan function gives only the principal value of the angle (i.e., modulo 2π).
- Formants are extracted by locating the peaks of the smoothed NGDF.

Figure 5.18 shows the NGDFs for the three vowels /a/, /u/, and /i/, where each curve has been smoothed by passing it through a 3-point Hanning window.

The choice of the LPC model's order determines the level and quantity of the spectral detail. If the model order is insufficient, then certain formants will not be adequately modelled into the spectrum, particularly in the case of closely spaced formants. Conversely, a model order which is excessive will deteriorate the signal-to-noise performance of the LPC-based spectral estimator. This is demonstrated by the presence of spurious spectral peaks from which it is then difficult to choose formant candidates.

Choosing a model order P as 18 has led to the presence of some spurious peaks in the NGDF of vowels. When reducing P to 14, the algorithm fails to resolve the case of having two closely spaced formants for some VRFs of the vowel /u/ (in these cases $F1$ and $F2$ are close to each other). Finally P has been set equal to 16, where adequate accuracy has been achieved. In the latter case, the problem of two closely spaced formants has been resolved, whereas some spurious peaks in the NGDF curve are to be eliminated by post-editing.

The input to the formant estimation algorithm is the vowel representative frames. The frames are supposed to be chosen in the vowel steady-state region. Thus, the NGDF's curve for such a VRF is expected to show prominent peaks relating to the vowel formants. In the post-editing, all peaks whose values are below 20% of the maximum value along the NGDF, are eliminated. The height of the peak in the NGDF can be shown to be inversely proportional to the bandwidth of the formant. Eliminating small peaks means eliminating peaks related to formants with wider bandwidths.

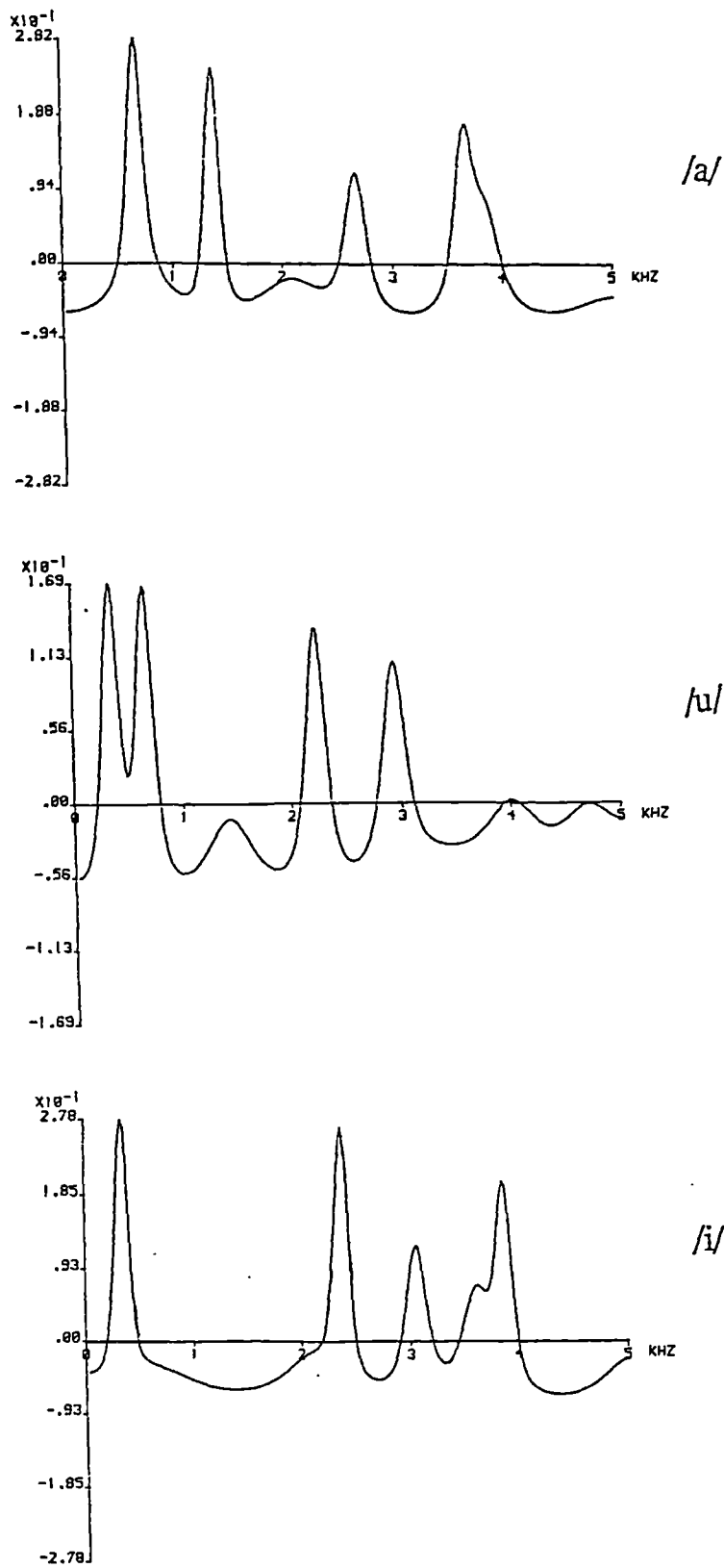


Fig. 5.18 The NGDFs of the three vowel types

5.4.3 Vowels in Pharyngealised Context

Arabic language has five pharyngealised consonants i.e., s, d, t, ṣ, and k. The effects of the pharyngealised consonant on the neighbouring vowels are demonstrated graphically in Chapter 3 by the speech spectrograms of the pharyngealised consonant-vowel pairs. It was found that in general, F1 and F2 along the vowel move closer to each other in pharyngealised context compared to the same vowel in plain (non-pharyngealised context).

After formant extraction, it is now possible to monitor the formant frequencies in plain and pharyngealised context. For example, the two words 'sabea' and 'ṣafaḥa' start with the consonant /s/ and its pharyngealised counterpart /ṣ/ respectively. The formant frequencies of the vowel representative frames of the first vowel in the syllables /sa/ and /ṣa/ are given as follows:

| | | | |
|------|-------------|--------------|---------------------|
| /sa/ | F1 = 605 Hz | F2 = 1445 Hz | $\Delta f = 840$ Hz |
| /ṣa/ | F1 = 586 Hz | F2 = 1035 Hz | $\Delta f = 449$ Hz |

other examples are as follows:

'kataba' , 'karaʔa' , 'kaana' , 'kaala' , 'niṣaam' , 'daraba'

| | | | |
|--------|-------------|--------------|----------------------|
| /ka/ | F1 = 625 Hz | F2 = 1738 Hz | $\Delta f = 1113$ Hz |
| /ḳa/ | F1 = 645 Hz | F2 = 1035 Hz | $\Delta f = 390$ Hz |
| /kaa/ | F1 = 684 Hz | F2 = 1602 Hz | $\Delta f = 918$ Hz |
| /ḳaa/ | F1 = 625 Hz | F2 = 977 Hz | $\Delta f = 352$ Hz |
| /ṣaa/ | F1 = 645 Hz | F2 = 957 Hz | $\Delta f = 312$ Hz |
| /ḍa/ | F1 = 605 Hz | F2 = 1035 Hz | $\Delta f = 430$ Hz |

The effect of the presence of a pharyngealised consonant on the formant frequencies of the following vowel is very clear. This pharyngealisation effect is not only confined to the neighbouring vowel, but it may extend over an entire word (to all vowels in the word) due to the coarticulation effect. The vowel of the syllable which contains a pharyngealised consonant in a certain word is definitely affected by the presence of the pharyngealised consonant, but other vowels in that word may be affected according to the place of the pharyngealised consonant within that word. For example, the word 'kaala' has the

syllabic structure /CVV-CV/, where the underlined consonant C refers to a pharyngealised consonant, and the formant frequencies of its vowels are:

| | | | |
|----------------|-------------|--------------|---------------------|
| / <u>k</u> aa/ | F1 = 625 Hz | F2 = 977 Hz | $\Delta f = 352$ Hz |
| /la/ | F1 = 684 Hz | F2 = 1543 Hz | $\Delta f = 859$ Hz |

It can be seen that the effect of the pharyngealisation did not extend to the second vowel in this word. Another example is the word 'ʔakaama' has the syllabic structure /CV-CVV-CV/, and the formant frequencies of its vowel are:

| | | | |
|----------------|-------------|--------------|---------------------|
| /na/ | F1 = 723 Hz | F2 = 1133 Hz | $\Delta f = 410$ Hz |
| / <u>k</u> aa/ | F1 = 586 Hz | F2 = 977 Hz | $\Delta f = 391$ Hz |
| /ma/ | F1 = 605 Hz | F2 = 1230 Hz | $\Delta f = 625$ Hz |

It is clear that the effect of pharyngealisation has extended over the first vowel and slightly spread to the third vowel. The word 'ʔawsaa' has the syllabic structure /CVC-CVV/ and the formant frequencies of its vowels are:

| | | | |
|----------------|-------------|--------------|---------------------|
| / <u>ʔ</u> aw/ | F1 = 645 Hz | F2 = 1094 Hz | $\Delta f = 349$ Hz |
| /saa/ | F1 = 586 Hz | F2 = 938 Hz | $\Delta f = 352$ Hz |

both vowels are affected in this word by the presence of the pharyngealised consonant /s/. It is worth to mention here that the consonant /r/ may be pharyngealised in some context, for example in the word 'ʔaraada' where the formant frequencies of its vowels are:

| | | | |
|-------|-------------|--------------|---------------------|
| /na/ | F1 = 605 Hz | F2 = 1339 Hz | $\Delta f = 734$ Hz |
| /raa/ | F1 = 686 Hz | F2 = 996 Hz | $\Delta f = 310$ Hz |
| /da/ | F1 = 605 Hz | F2 = 1387 Hz | $\Delta f = 782$ Hz |

in this case, only the vowel following the pharyngealised consonant is affected.

As a result, in the pattern /CVCV/, the pharyngealisation effect is confined to the vowel following the pharyngealised consonant, and does not extend to the other vowel in the word. In the patterns /CVCV/, /CVCCV/, and /CVCCV/, the pharyngealisation effect extends over the vowel preceding the pharyngealised consonant to the vowel following

it.

The previous study concerns the vowels /a/ and /aa/. Actually we could not extend this study to the other vowels (i.e., /u/, /uu/, /i/, and /ii/). This is because the changes in the formant frequencies of a vowel near to a pharyngealised consonant compared to their values for a vowel near to a plain consonant, is not as noticeable as in the case of the vowels /a/ and /aa/. Such differences do exist, but they are not large enough to be distinguishable over different realisations of those vowels in different contexts.

Returning to the results for the vowels /a/ and /aa/, it can be concluded that detecting a pharyngealised vowel (according to its formant frequencies) leads to the prediction of the relative location of the pharyngealised consonant in that word. Also any syllable which contains a pharyngealised vowel can be labelled as a pharyngealised syllable.

5.5 Summary

In this chapter, two main procedures have been introduced, i.e., voiced-unvoiced-silence segmentation and vowel recognition.

The first procedure segments the speech signal into voiced speech segments, unvoiced speech segments and silence (no speech) segments. Parameters such as energy, zero-crossing rate, autocorrelation coefficients, and some parameters from the pitch detection algorithm have been employed in this procedure. Smoothing and editing have been carried out on the V-UV-S contour of a word in order to reach final accurate segmentations. The accuracy of this preliminary segmentation must be very high, because errors made at this level will spread to the final segmentation process and might affect the final recognition accuracy.

The vowel recognition procedure consists of two main parts, i.e., the vowel detection phase and the vowel identification phase. In the vowel detection phase, the peaks of the energy contour are used to locate the vowel steady-state regions, where vowel representative frames (VRFs) are chosen to be used in the vowel identification phase. Heavy smoothing and editing are performed on the energy contour in order to eliminate false and spurious peaks. Durational information is used at this level to distinguish between short and long vowels. The concept of the vowel estimated duration (VED) has been introduced in this chapter. The VRF and the VED are going to be employed in the

final segmentation and error correcting processes.

In the vowel identification phase, two techniques have been implemented to identify the vowel type, i.e., the vector quantisation (VQ) method and the formant method. The implemented VQ method is based on the full search binary-split algorithm. The VQ method is actually based on pattern matching approach, and gave high recognition accuracy. The formant method is based on the extraction of the first two formants for each vowel and uses them for vowel identification through a decision tree algorithm. The formant method gives more freedom than the vector quantisation method, especially for multi-speaker systems. In a multi-speaker system, it is preferable to adopt a formant normalisation algorithm to account for speaker variations, or to use some sort of speaker adaptation by asking each new speaker to utter a few words from which the system can adjust the boundaries between the vowel regions in the F1-F2 plane. Also the formant method gives extra information about vowels in pharyngealised context. Formant frequencies are extracted from the derivatives of the phase spectra (or the negative of the group delay function) of the vowel representative frame.

The accuracy of the V-UV-S segmentation is almost perfect. Error occurs in the form of adding a short segment, or extending the segment at the end of a word, due to the presence of breathing noise. This error is tackled in the editing process, and could also be avoided by adjusting the position of the microphone and training the speaker to avoid as much as possible generating such noise.

The accuracy of the vowel detection phase is about 99%. The remaining 1% errors are unavoidable in this algorithm, since they occur when low intensity vowels appear adjacent to relatively high intensity voiced consonants. The accuracy of the vowel identification phase is also about 99% for the VQ method (for the used speech database). The accuracy of the formant method is perfect. The formant method also gave perfect accuracy for female speaker, after modifying the boundaries between the three vowel regions in the F1-F2 plane. The 1% errors made in the vowel recognition procedure are tackled in the final error correcting procedure described in Chapter 7.

Finally, recalling some statistical results from Chapter 3, vowels represent about 43% of the total number of phonemes in the lexical database (10,000 words containing 75875 vowels). The vowel /a/ represents about 47% of the total number of vowels, and /aa/ about 13%, i.e. both of them represent 60%. The other four vowels (i.e., /u/, /uu/, /i/,

and /ii/) represent 40%. Thus, having vowels equal to 43% of the total number of phonemes makes the vowel recognition task very vital in any large vocabulary Arabic speech recognition system. Improving the accuracy of the vowel recognition algorithm will surely improve the accuracy of the recognition system.

The following chapter presents the spectral transition detection stage in the speech recognition model of Figure 4.5. The results of this stage are used in the segmentation and labelling processes presented in Chapter 7.

Chapter 6

Spectral Variation Contour and its Application to Speech Segmentation

6.1 Introduction

Speech segmentation is often defined to be the process of dividing the speech waveform into a series of discrete acoustic states which are related to a phonemic transcription of the utterance (an utterance is made of a concatenation of several phonemes).

Speech consists of sustained sound segments where the acoustic characteristics of the sound are similar, and of transitional sound segments, where the acoustic characteristics vary with time. Thus the spectrum of an utterance is composed of alternating steady-state and transition regions.

In the previous chapter, the steady-state regions for vowels, where the labelling is most reliable, have been located through the use of the energy contour. In this chapter, an automatic method for detecting the boundaries between adjacent phonemes, including vowels, is introduced. The outcome of this method along with the result of the preliminary V-UV-S segmentation and vowel recognition, are all employed for word segmentation in the next chapter.

For the purpose of segmentation, the spectral variation along a word is to be extracted from the speech signal. Conventional methods for extracting spectral movement information are mainly based on formant trajectory estimation. However, tracking formant trajectories is usually difficult and error-prone. Therefore, it is desirable to extract spectral variation without resorting to formant tracking.

An alternative method is the extraction of a spectral variation contour along a word. This contour should manifest the transition between sub-word units (these units should be single phonemes).

6.2 Spectral Variation Contour

The spectral variation function represents a time-varying signal which is evaluated at equally spaced points of time. The speech signal of a word can be divided into a sequence of equally spaced frames. Each frame is represented by an N-dimensional vector. This vector represents the spectral envelope, which is a close approximation of the vocal tract transfer function. The sequence of vectors represent a curve in an N-dimensional space. The spectral variation function is calculated at each point (frame) of this curve as the average spectral distance between this point and the neighbouring points (frames) which lie within a window of $2L+1$ points. This function is defined as:

$$sv(n) = \frac{1}{2L} \sum_{j=-L}^L d(V_n, V_{n+j}) \quad (6.1)$$

where $sv(n)$ is the value of the spectral variation function at frame n , V_n and V_{n+j} are the spectral vectors (parametric representation of the spectrum) of frames n and $n+j$ respectively, j is in the range $[-L, L]$, and d is the distortion measure between a pair of spectral vectors. Each speech frame is represented by an N-dimensional vector as:

$$V = \{ C(1), C(2), \dots, C(N) \} \quad (6.2)$$

where $C(i)$ is the i th parameter. The distance d is the Euclidian distance between two vectors a and b and defined as:

$$d(V_a, V_b) = \left[\sum_{i=1}^N (C_a(i) - C_b(i))^2 \right]^{1/2} \quad (6.3)$$

For the purpose of spectral variation function calculation, the mel-frequency cepstral coefficients are employed as a parametric representation of the speech signal.

The sequence $[sv(n)]$ along a word is called the spectral variation contour (SV contour). This contour is used to extract the transitional information which is associated with the phonemic boundaries as will be shown later on.

6.3 Parametric Representation

The selection of the best parametric representation of the acoustic data is an important task in the design of any speech recognition system. In Section 2.3.1, several parametric representations of the speech signal have been presented, such as LPC parameters, filter bank parameters, and cepstral parameters. The objectives of these parametric representations are to compress the speech data by eliminating information not pertinent to the phonetic analysis of the data and to enhance those aspects of the signal that contribute significantly to the detection of phonetic differences. Cepstrum parameters are chosen in this study as a spectral representation of the speech signal. The cepstrum parameters provide a compact (low dimensional) representation of the vocal tract transfer function. They also allow the use of a simple distance measure in the computation of the spectral variation contour.

6.3.1 Cepstrum Parameters

According to the simplified model of speech production (see Figure 2.1), the speech signal $s(n)$ is given as the convolution between the source signal $g(n)$ and the vocal tract impulse response $h(n)$ (for a short-time signal) as:

$$s(n) = g(n) * h(n) \quad (6.4)$$

In the Z-domain Eq. (6.4) is written as:

$$S(z) = G(z) H(z) \quad (6.5)$$

where $S(z)$, $G(z)$, and $H(z)$ are the transfer functions of the speech signal, the source signal, and the vocal tract model (filter), respectively. The transfer function of the vocal tract takes different shapes when pronouncing different phonemes. The source transfer function is more influenced by higher-level linguistic phenomena rather than by phonemics. One way of resolving the convolution in Eq. (6.4) and the multiplication in (6.5) is to use the logarithm in the frequency or the Z-domain. Thus Eq. (6.5) becomes:

$$\log (S(z)) = \log (G(z)) + \log (H(z)) \quad (6.6)$$

where the multiplication is converted to addition. If $\log(S(z))$ is converted back to the time domain, the resultant signal is drastically different from the original time-domain signal. A sequence is obtained in which the impulse response and the source signal are superimposed in an additive way, and are therefore expected to be much more easily separated than in the original signal.

For speech signals, and especially for signal frames (short-time signals), Eq. (6.6) is valid when evaluated at the unit circle, i.e., for $z=e^{j\omega T}$, and it can be rewritten in terms of the discrete Fourier transform (DFT) as:

$$\log(S(m)) = \log(G(m)) + \log(H(m)) \quad (6.7)$$

where $S(m)$ is the discrete spectrum at sample m in the frequency domain. The inverse Fourier transform of $\log(S(m))$ is called the complex cepstrum $X_s(n)$, which is a special case of the homomorphic processing [33]. For speech processing, further simplification is possible, where Eq. (6.7) is likewise valid for the power or the amplitude spectrum of the speech as:

$$\log|S(m)| = \log|G(m)| + \log|H(m)| \quad (6.8)$$

Taking the inverse DFT of Eq. (6.8) yields:

$$C_s(n) = C_g(n) + C_h(n) \quad (6.9)$$

where $C_s(n)$ is called the power cepstrum or simply the cepstrum. The cepstrum $C_s(n)$ is equal to the even part (real part) of the complex cepstrum $X_s(n)$.

The cepstral components related to the vocal tract are selected by the cepstrum window $l(n)$ which is of the form:

$$l(n) = \begin{cases} 1, & |n| < N \\ 0, & |n| \geq N \end{cases} \quad (6.10)$$

where N is chosen to be less than the pitch (fundamental frequency F_0) period. Also, F_0 can be extracted from $C_s(n)$ by remembering that the $G(m)$ of a voiced signal is a pulse train, and it is transferred to the cepstrum domain into another pulse train with no inherent information other than its periodicity [128].

The coefficients of Eq. (6.9) are called linear frequency cepstral coefficients (LFCCs). The LFCCs for a frame of $2K$ speech samples are calculated according to the following steps:

- Hamming windowing ($2K$ points).
- FFT Computation ($2K$ points).
- Log-magnitude of the DFT coefficient (K points).
- Cosine transform of the K log-magnitude components (inverse DFT).

Since the magnitude spectrum is a real even function, the inverse DFT can be achieved by the cosine transform (real Fourier transform). Thus the i th component is given as:

$$\text{LFCC}_i = \sum_{k=0}^K Y_k \cos\left(\frac{\pi i k}{K}\right) \quad (6.11)$$

where Y_k is the log-magnitude of the k th DFT coefficient, and K is the number of DFT coefficients within half the sampling frequency range. The first N LFCCs are used to represent the vocal tract, where N is less than the minimum expected pitch period.

6.3.2 Mel-Frequency Cepstral Parameters

The cochlea in the human inner ear is thought to perform a continuous broad-band analysis of the sound which enters the ear, and transmits the results to the brain. The frequency range over which the human ear is able to perceive sounds is often divided according to the concept of critical bands (see appendix C). A critical band can be viewed as a bandpass filter whose frequency response corresponds roughly to the tuning curves of auditory neurons. Thus, the linear frequency scale can be warped to follow either the Bark scale (one Bark unit covers one critical bandwidth), or to follow the mel scale (the mel is the unit of pitch, where one bark corresponds roughly to a pitch interval of 100 mels). The mel scale is essentially linear at low frequencies below 1KHz and logarithmic at higher frequencies above 1KHz [102]; see Appendix C for more details.

The mel-frequency cepstral coefficients (MFCCs) result from replacing the linear frequency scale by a mel scale, where the frequency range (of the input signal) is divided into a bank of bandpass filters. These filters are linearly spaced at low frequencies and logarithmically at high frequencies. The MFCCs are computed as the result of a cosine transform of the logarithm of the short-time energy spectrum of the filter bank outputs. Thus the i th MFCC is given as:

$$\text{MFCC}_i = \sum_{k=1}^K E_k \cos \left[i \left(k - \frac{1}{2} \right) \frac{\pi}{K} \right] \quad (6.12)$$

where $i=1,2,\dots,N$ and $k=1,2,\dots,K$. N is the number of required MFCCs and K is the number of filters in the filter bank which cover the frequency range of the input signal. E_k represents the logarithm of the energy output of the k th filter.

For the computation of MFCCs, a bank of 22 triangular bandpass filters have been simulated as shown in Figure 6.1 [35]. Table 6.1 displays the centre frequencies and bandwidth of the filters in the filter bank [140]. This bank of bandpass filters separates the frequency spectrum of interest into various frequency bands according to the mel scale (see appendix C for more details). The spacing of the filters is implemented in such a way that they are continuous over the frequency spectrum and the composite spectrum (transfer function) of the overall filter bank is essentially flat, i.e. no sharp valleys between adjacent filters.

Returning to Eq. (6.12), the coefficient MFCC_0 represents the average energy in the speech frame and is discarded as a form of amplitude normalisation. This can be explained by normalising the log-energy of each channel by MFCC_0 . Thus, the normalised 0th coefficient becomes equal to zero, and the normalised log-energy of each channel becomes equal to $(E_k - \text{MFCC}_0)$. Substituting the latter value in Eq. (6.12) yields exactly the same equation, where the cosine transform of a fixed value (MFCC_0) is zero.

As in the case of the linear frequency cepstral coefficients, the first N mel-frequency cepstral coefficients are associated with the impulse response of the vocal tract.

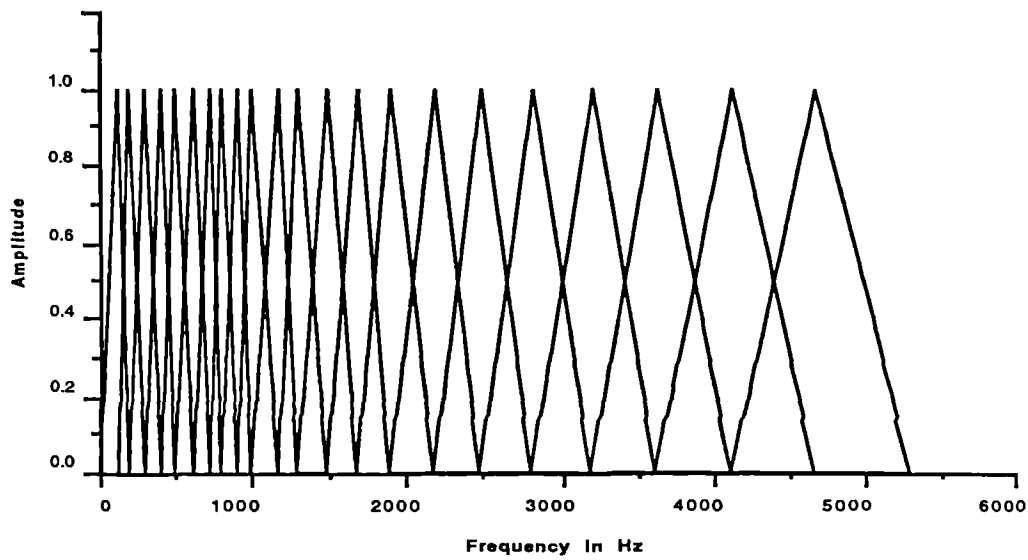


Fig. 6.1 Bank of 22 tringular filters

| Filter No. | Centre Frequency Hz | Bandwidth Hz |
|------------|------------------------|-----------------|
| 1 | 100 | 100 |
| 2 | 200 | 100 |
| 3 | 300 | 100 |
| 4 | 400 | 100 |
| 5 | 500 | 100 |
| 6 | 600 | 100 |
| 7 | 700 | 100 |
| 8 | 800 | 100 |
| 9 | 900 | 100 |
| 10 | 1000 | 118 |
| 11 | 1137 | 146 |
| 12 | 1292 | 166 |
| 13 | 1469 | 189 |
| 14 | 1671 | 215 |
| 15 | 1899 | 244 |
| 16 | 2159 | 278 |
| 17 | 2455 | 316 |
| 18 | 2791 | 359 |
| 19 | 3173 | 408 |
| 20 | 3607 | 464 |
| 21 | 4101 | 527 |
| 22 | 4662 | 599 |

Table 6.1 Filter bank centre frequencies and bandwidths (mel scale)

In some speech recognition experiments based on the template matching approach [35, 141], it has been reported that using the first six MFCCs gives better recognition accuracy than any other set of the following parameters: DFT coefficients, LFCCs, filter bank parameters, LPC parameters, and LPC-derived cepstral parameters (LPCCs) (see Section 2.3.1).

These results have been confirmed by our own experience, where at the beginning of this research work, a recognition experiment based on *template matching (using DTW)* has been conducted. In this experiment, two tests have been performed. In the first test, a set of 50 English words (comprising the alpha-digits and a few other words) uttered by three native speakers, has been used as a speech database. In the second test, 10 Arabic words (the 10 digits) uttered 10 times by the one speaker, have been used as a database. Both tests have shown that six MFCCs give better accuracy than 16-LPC parameters. These results indicate superior performance of the MFCC when compared with other parametric representations, and the first six MFCCs succeed in capturing the significant acoustic information. This compact representation is also more successful than other parametric representations in indicating the phonetic significance of the difference between a pair of spectra by computing a distortion or distance measure between their representative vectors.

6.3.3 MFCCs Computation

As stated in Chapter 5, the speech signal is band limited to the range 60-4800 Hz, sampled at 10 KHz, and coded with 12 bits.

Figure 6.2 shows a block diagram for the processes involved in computing the MFCCs. In this diagram, a vector of N -MFCC is computed for each frame of 6.4 msec length using a block of 25.6 msec with 75% overlapping for each analysis block. A spectral analysis is performed through a bank of K filters, to yield a K -dimensional energy vector. Then, a cosine transform according to Eq. (6.12) is performed, and a vector of N MFCCs is given for each frame. K is taken equal to 22, and N is chosen equal to 6.

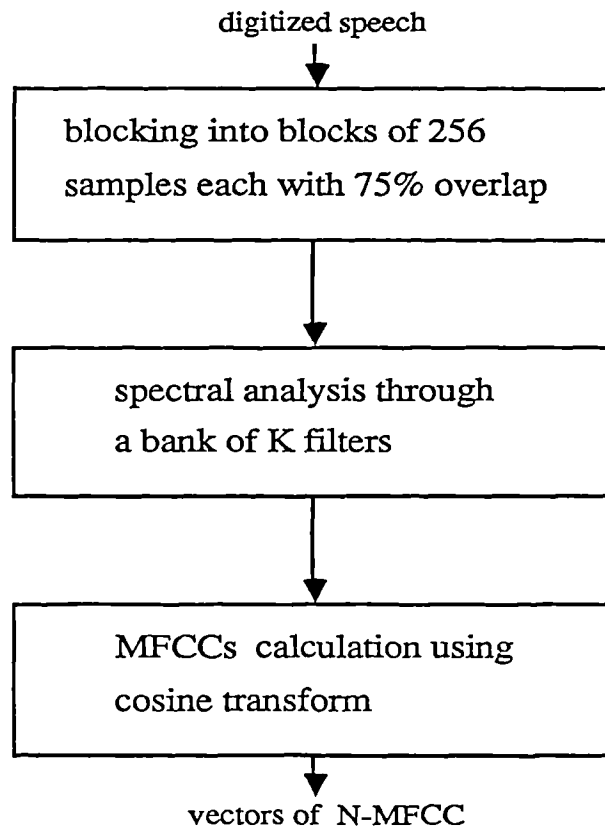


Fig. 6.2 MFCCs Computation

a) Filter Bank

The spectral analysis has been achieved by using an FFT-based Filter bank. Figure 6.3 displays a block diagram for computing the K-dimensional log-energy vector. The spectral analysis for each block is performed by a short-time discrete Fourier transform (DFT) according to following equation:

$$X(m) = \sum_{i=0}^{M} x(i) e^{-j \frac{2\pi}{M} im} \quad (6.13)$$

where $x(i)$, $i=1,2,\dots,M$ are the speech samples in the analysis block. $X(m)$ is the m th component (sample) in the frequency domain, where $m=1,2,\dots,M$. The DFT can be

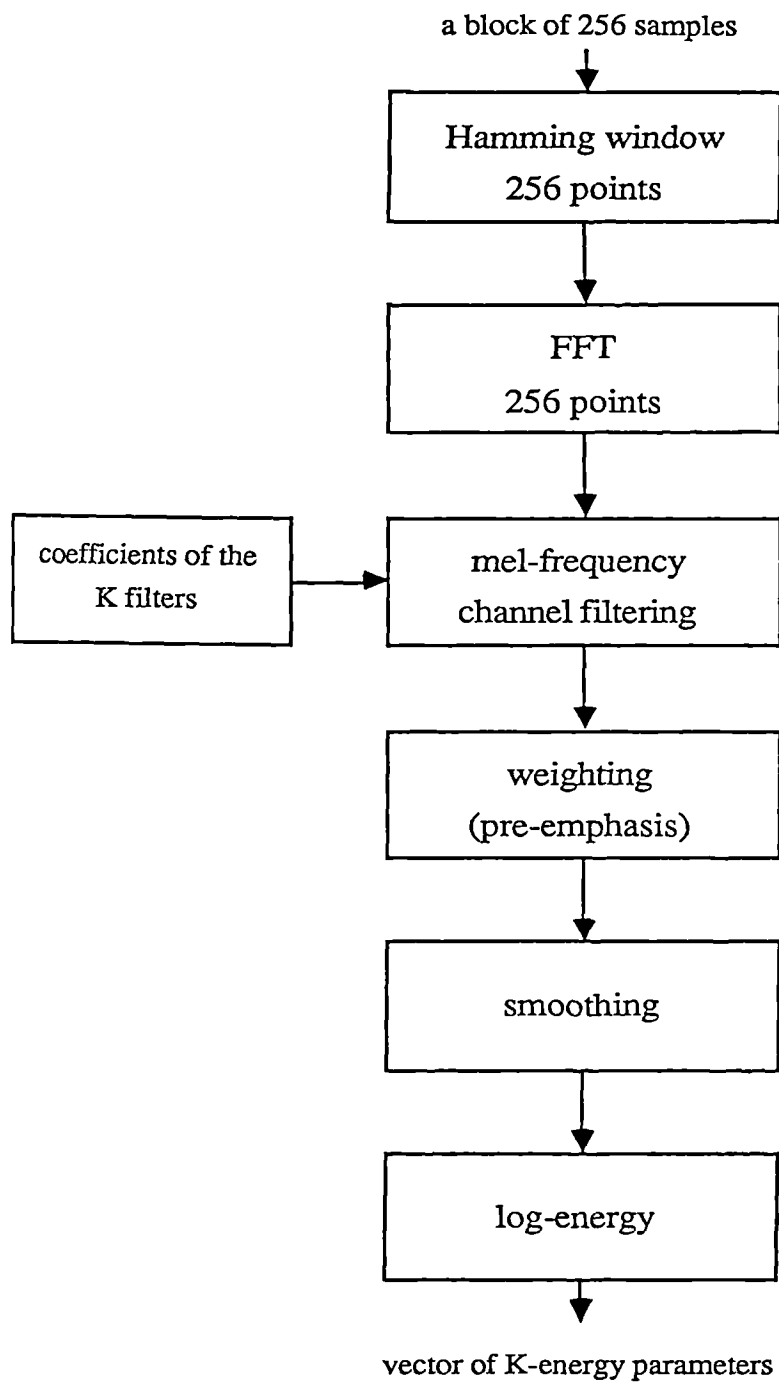


Fig. 6.3 Mel-frequency channel filtering

efficiently computed by an FFT algorithm. Hamming windowing is applied to the speech block before the DFT, in order to reduce the noise in the spectrum which would have occurred if the block was left without Hamming windowing (the rectangular window has relatively higher side lobes in its spectrum compared with the Hamming window, and produces a ragged or noisy spectrum). The Hamming window is given as:

$$HW(i) = 0.54 - 0.45 \cos\left(\frac{2\pi i}{M}\right) \quad (6.14)$$

where $0 \leq i \leq M-1$. Then the magnitude ($|X(m)|$) of the first 128 DFT coefficients, which represent the spectral components in the range 0-5000 Hz, is computed.

The coefficients of the 22 triangular filters of Figure 6.1 are calculated as follows. From Figure 6.4, the coefficients of a filter k are given as:

coefficients at the leading edge of the filter:

$$a_k(m) = \frac{F_m - F_1}{F_2 - F_1} \quad F_1 \leq F_m \leq F_2 \quad (6.15)$$

coefficients at the trailing edge of the filter:

$$a_k(m) = \frac{F_3 - F_m}{F_3 - F_2} \quad F_2 < F_m \leq F_3 \quad (6.16)$$

where F_m is the frequency at the m th component which is given as:

$$F_m = \frac{\text{sampling frequency}}{\text{number of DFT points}} m = \frac{10000}{256} m \quad (6.17)$$

where $1 \leq m \leq 128$. The resolution or the spacing between the frequency components is about 40 Hz. Thus, the coefficients of the 22 filters given in Figure 6.1 are computed in the same way, and each filter has a number of coefficients (non zero value) according to its bandwidth.

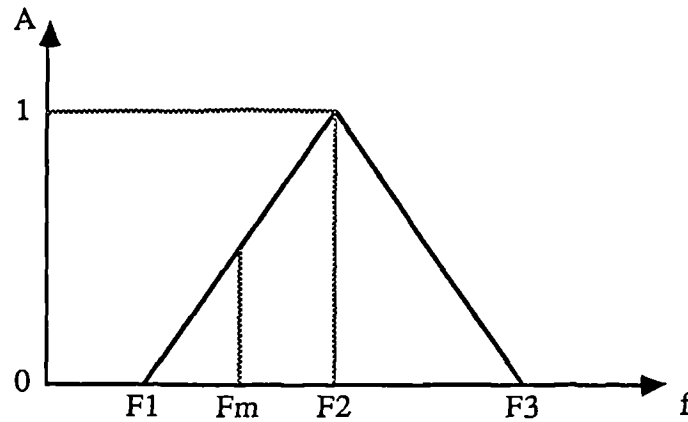


Fig. 6.4 Triangular filter

In general, the energy at the output of a filter k is given as follows:

$$E(k) = W(k) \sum_{m=1}^{M/2} (a_k(m) |X(m)|)^2 \quad (6.18)$$

where $a_k(m)$ is equal to zero for m outside the frequency range of the k th filter. $W(k)$ is a weighting function. The energies at the filter outputs are weighted by an equal-loudness curve. This fixed curve approximates the auditory response over a mid-range intensity level (it is actually a scaled version of the inverse of the hearing sensitivity curve; see Appendix C). This curve has a slope of +10 dB/oct in the range of 0.1-0.4 KHz, flat in the range of 0.4-1.2 KHz, +6 dB/oct in the range of 1.2-3.1 KHz, and flat in the range of 3.1-5 KHz [142]. This weighting process is almost equivalent to the pre-emphasis process. Thus, the log-energy E_k is computed as the logarithm of $E(k)$ given in Eq. (6.18). The final result is a vector of K parameters for each speech frame.

In general, the energy contours will fluctuate very rapidly depending on the exact details of the speech waveform (i.e., depending on the placement of the analysis window and the fraction of pitch periods within the window). Linear smoothing is carried out on the energy contour over time for each channel in order to remove fluctuations over time which result from the short-time analysis of the speech signal. Linear smoothing is achieved by using a 5-point Hanning window, which has the following impulse response:

$$\begin{array}{ll}
 w_h(n) = 3/9 & n = 0 \\
 & |n| = 1 \\
 & |n| = 2 \\
 & |n| \geq 3
 \end{array} \quad (6.19)$$

Smoothing is actually applied to the linear energy contour of each channel ($E(k)$).

b) Cepstral Parameters

Six mel-frequency cepstral parameters are calculated according to Eq. (6.12) by employing the vector of K-log-energy parameters. Thus each speech frame j (6.4 msec) is described by a vector of six MFCCs as follows:

$$V_j = \{ C_j(1), C_j(2), C_j(3), C_j(4), C_j(5), C_j(6) \} \quad (6.20)$$

The time variation of the V_j vectors defines a pattern in a 6-dimensional space for a given word. The resultant cepstral parameters are also smoothed over time using a 3-point Hanning window followed by a 5-point window. The smoothing is performed on each parameter $C(i)$ over time (where i is in the range of 1-6).

6.4 Transition Detection and Segmentation

The spectral variation (SV) contour given in equation (6.1) is calculated over each word as the average spectral distortion between a frame n and the neighbouring frames within a window of $2L+1$ frames. The resulting SV contour is smoothed through a 3-point Hanning window to remove unwanted fluctuation along this contour.

Figure 6.5 shows graphs for the word 'markazu', where a) represents the speech spectrogram for this word, b) the speech signal, c) the V-UV-S contour, d) the energy ES contour, and e) the SV contour. It can be seen from these figures that regions with high $sv(n)$ values are usually associated with transient sounds, while regions with low $sv(n)$ values are usually associated with steady-state sounds. By comparing the SV contour with the speech spectrogram of this word, it can be clearly seen that the peaks of the SV contour are situated within the transitional region between two adjacent

phonemes. These peak points can be taken as rough estimates of the phoneme boundaries. A simple peak-picking algorithm can be employed to detect these peak points along the SV contour. Figure 6.5e indicates that this word has seven segments or sub-word units which in this case represent the phonemes of the word 'markazu'. The distance between any two successive peak points represents the length or the duration of that segment (or phoneme). As a result, information from the SV contour can be employed to automatically segment the speech signal into phonemic units regardless of their phonemic identity.

By combining information from the V-UV-S and ES contour with the results of the vowel identification procedure and the SV contour, it can be concluded that the word 'markazu' consists of seven phonemes. Three of these phonemes are the vowels /a/, /a/, and /u/, and the other four are consonants. The syllabic pattern of this word is /CVC-CV-CV/.

The segmentation process is actually more complicated than in the above case, and it is explained in detail in the next chapter. In the rest of this chapter, it is demonstrated how transition information can be reliably extracted from SV contours.

In calculating the window of the $sv(n)$ function, it is empirically found that $L=2$ ($2L+1 = 5$ frames or 32.5 msec) gives better results than other values. It is noticed that some of the SV contours' peaks are removed (or smoothed) when L is taken greater than 2, especially those peaks associated with the transition between two voiced phonemes where one of them may have a short duration (as low as 6 frames).

In fact, the SV contour of Figure 6.5e is very clean (i.e., not fuzzy), but this is not always the case for all words. Two major problems are confronted in this respect, i.e.,

- detecting spurious peaks along the SV contour
- missing some peaks which are related to the transition between phonemes

False or spurious peaks along SV contour are eliminated via two steps. The first step is to set a transition threshold (SV_{th}), and to discard all peaks whose $sv(n)$ values fall below this threshold. The second step is to tackle the false peaks which pass the first test and

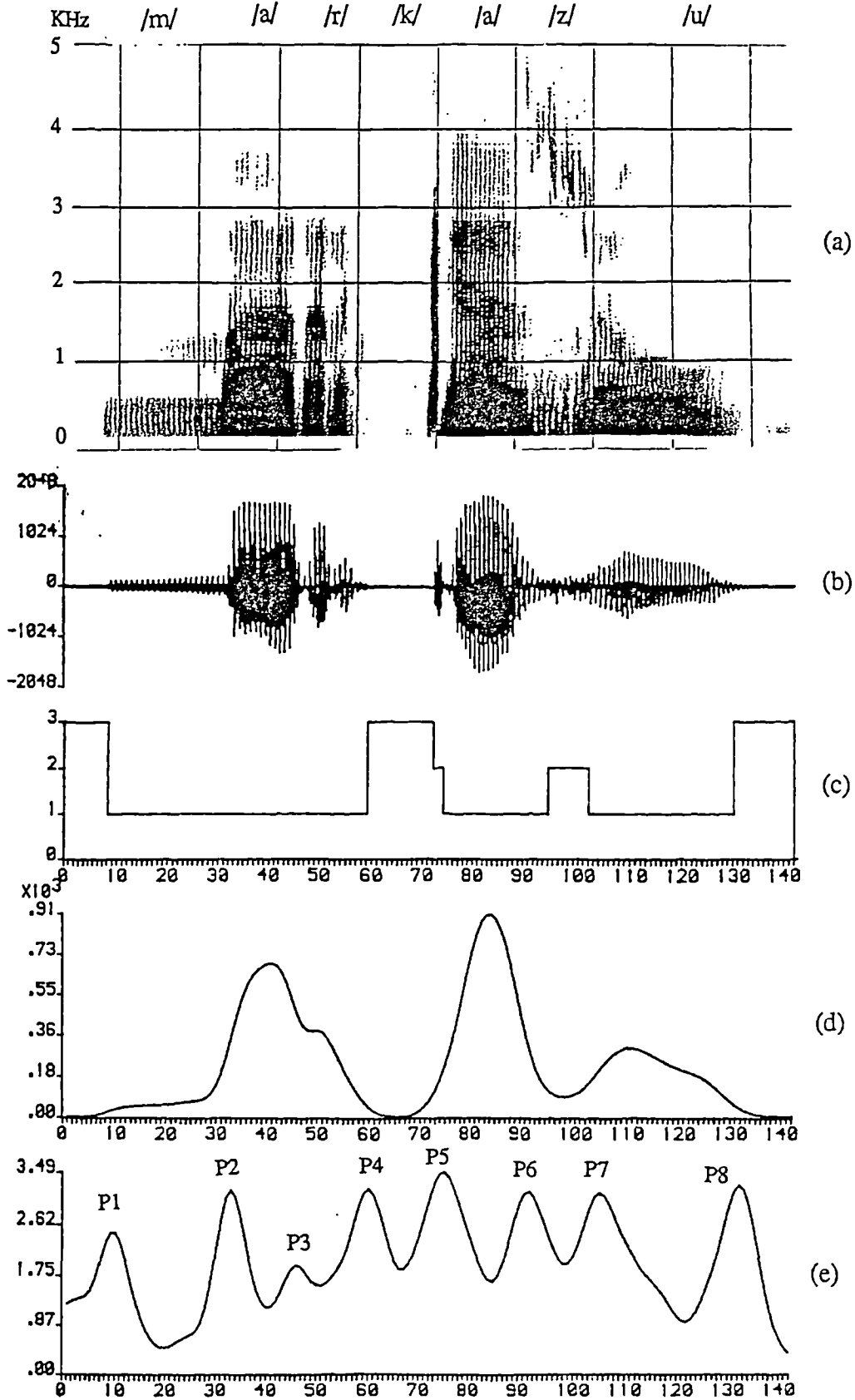


Fig. 6.5 Graphs for the word 'markazu'
 a) the speech spectrogram, b) the speech signal
 c) the V-UV-S contour, d) the ES contour, e) the SV contour

are taken as genuine peaks by using phonological constraints. The latter case, and the problem of missing peaks, are handled by a special procedure which corrects the segmentation error, as will be explained in the next chapter.

A self-normalised transition threshold SVth is calculated from the function $sv(n)$ of each word. The SVth has been chosen as equal to the mean value of the $sv(n)$ function along a word. It has been noticed from the observation of many SV contours for many words, that, a constant threshold may be used for most of the words. This is because of the use of mel-frequency cepstral parameters, which are considered as normalised parameters and less sensitive to the voice level. However, the mean value of $sv(n)$ over a word is adopted as a transitional threshold. The use of this threshold has led almost to correct results for all the test words in the speech database. It is believed that there are some relations between the value of the transitional threshold SVth on one hand, and the mean value of $sv(n)$ along the sv contour, the maximum value, and the duration information (such as the speaking rate and the length of the relevant phonemes) on the other hand.

Figure 6.6 shows graphs for the word 'θamaanija', where a) shows the speech spectrogram, b) the speech signal, c) the V-UV-S contour, d) the ES contour, and e) the SV contour. The SV contour displays 7 prominent peaks (their values are above SVth), which implies that it has 6 segments. These peaks are located at the following frame numbers along the word:

| | | | | | | | |
|-------|----|----|----|-----|-----|-----|-----|
| peak | P1 | P2 | P3 | P4 | P5 | P6 | P7 |
| frame | 37 | 50 | 65 | 103 | 112 | 138 | 154 |

These peaks indicate the phoneme boundaries. The segments associated with the phonemes of this word are given as follows:

| | | | |
|-----------|---------|------------------|-----------|
| segment 1 | P1 - P2 | for the phoneme | /a/ |
| segment 2 | P2 - P3 | for the phoneme | /m/ |
| segment 3 | P3 - P4 | for the phoneme | /aa/ |
| segment 4 | P4 - P5 | for the phoneme | /n/ |
| segment 5 | P5 - P6 | for the phonemes | /i/ - /j/ |
| segment 6 | P6 - P7 | for the phoneme | /a/ |

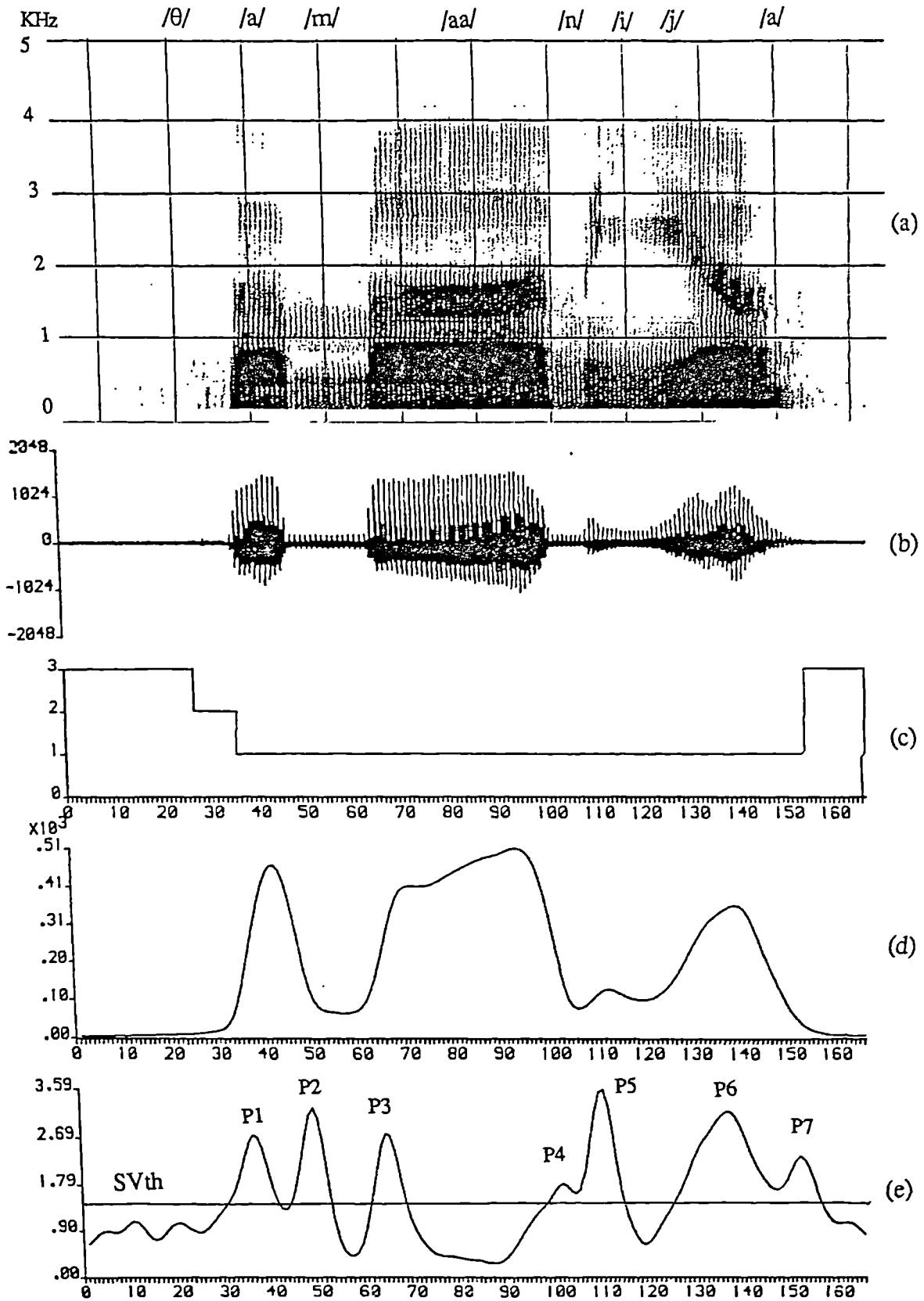


Fig. 6.6 Graphs for the word 'θamaanija'
 a) the speech spectrogram, b) the speech signal
 c) the V-UV-S contour, d) the ES contour, e) the SV contour

This word comprises 8 phonemes, four vowels and four consonants. The first phoneme /θ/ is unvoiced, while the rest are all voiced. It can be seen from the above information that two peaks are missing along the SV contour. The first peak should be at the beginning of the word and the second peak should be between P5 and P6 to indicate the border between the phonemes /i/ and /j/.

The first missing peak at the beginning of the word should be at frame 28 to mark the boundary between the silence and the phoneme /θ/. This peak did not occur because /θ/ is a weak unvoiced fricative and was hardly detected through the weak fricative detection algorithm of Section 5.2.4. Thus, this peak has been recovered from the V-UV-S contour which determines the word's endpoints.

By looking at the spectrogram in Figure 6.6a, it can be seen that there is no major spectral transition or discontinuity between the vowel /i/ and the semivowel /j/, since both of them have almost the same formant structure (of course in the steady-state region). For this reason, the SV function was not able to detect any transition between these two phonemes. However, this problem can be solved by comparing the estimated duration of the vowel /i/ (which results from the vowel detection procedure) with the length of segment 5 (in the range P5-P6) which contains this vowel. As a result of this comparison, it can be assumed that a voiced consonant with the same spectral structure as the previous vowel should be present after the vowel within the boundaries P5-P6. This case and others are dealt with later, in Chapter 7.

In Figure 6.6e, the peak P4 at frame 103 would be smoothed out if the $sv(n)$ calculation window $(2L+1)$ was larger than 5 frames. Sometimes such peaks might be removed because of the smoothing process which is applied to the energy and MFCC parameters. Smoothing is very important to obtain a clean SV contour, but it may cause some errors, because it removes some weak transitional peaks along the SV contour.

Figure 6.7 shows graphs for the word 'safaha', where a) shows the speech signal, b) the V-UV-S contour, c) the ES contour, and d) the SV contour. This word contains three vowels (three /a/) and three unvoiced consonants (/s/, /f/, /h/). It can be seen from figures 6.5e, 6.6e, and 6.7d that these SV contours show prominent peaks at the transition between voiced and unvoiced phonemes. This type of transition always leads to prominent peaks along the SV contours, as can be seen clearly from the spectrograms of the consonant-vowel pairs given in Chapter 3 and Appendix B.

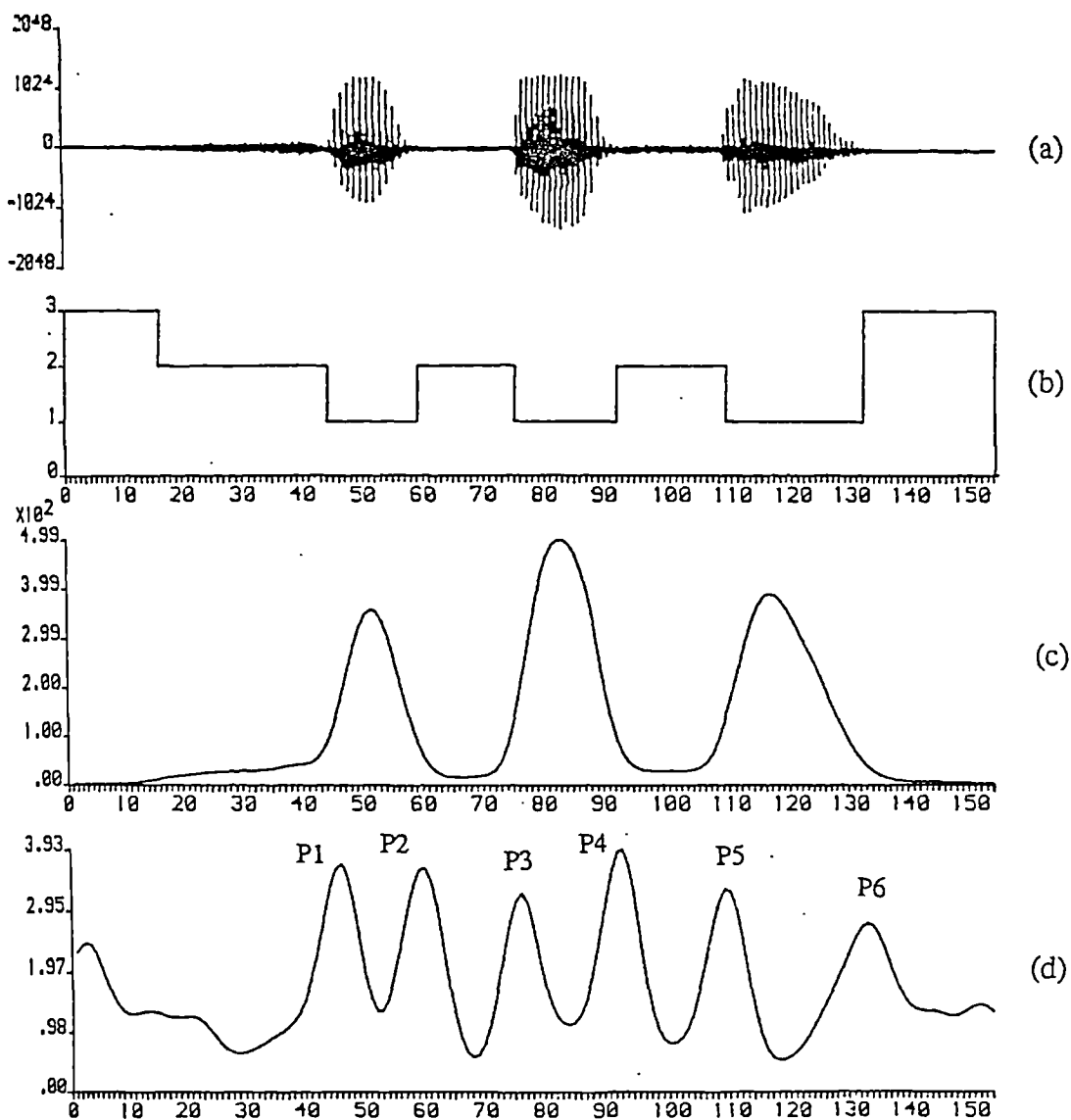


Fig. 6.7 Graphs for the word 'safaha'
 a) the speech signal, b) the V-UV-S contour
 c) the ES contour, d) the SV contour

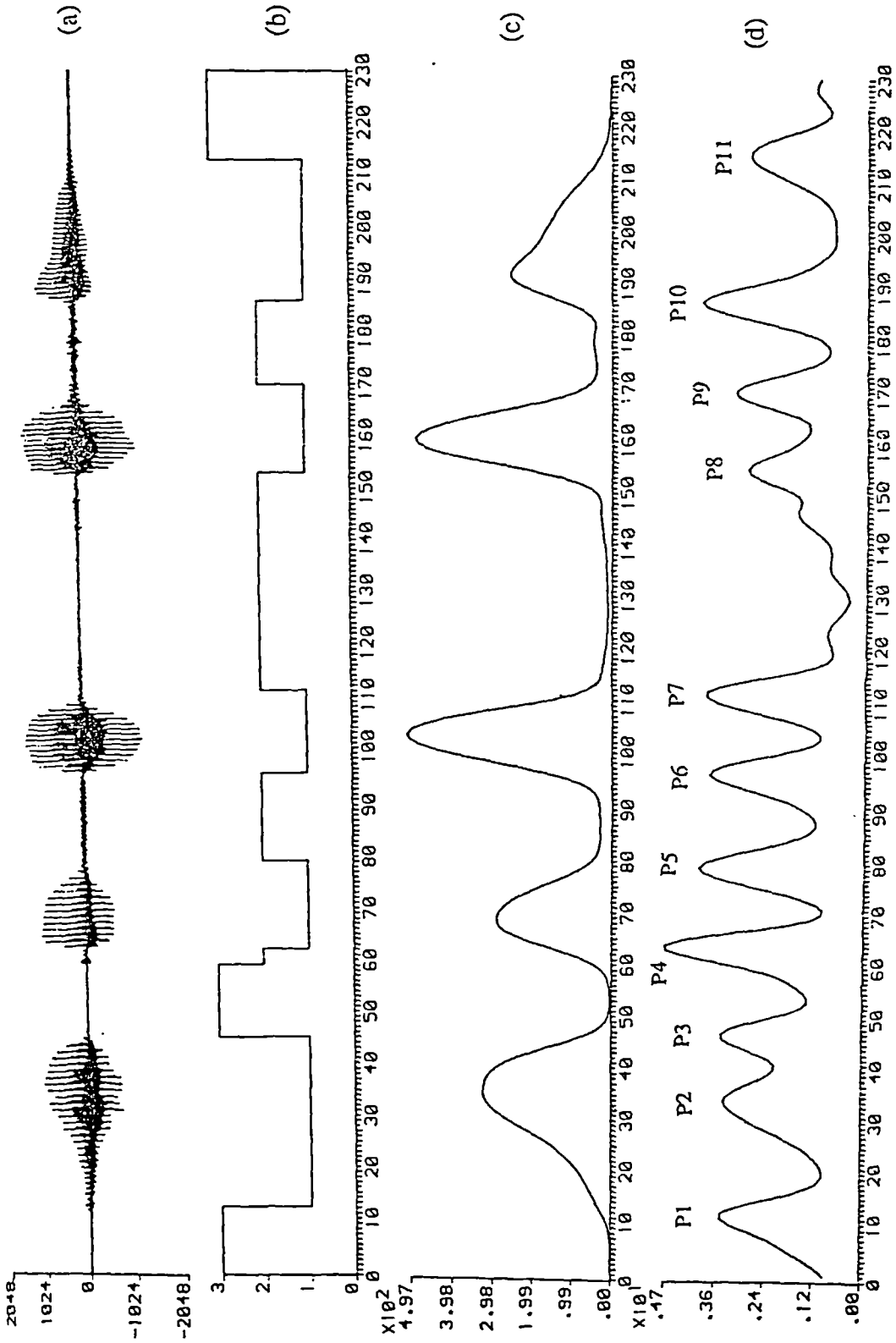


Fig. 6.8 Graphs for the word 'jatasaffathu'

a) the speech signal, b) the V-UV-S contour, c) the ES contour, d) the SV contour

Figure 6.8 shows graphs for the word 'jataṣaffaḥu', where a) shows the speech signal, b) the V-UV-S contour, c) the ES contour, and d) the SV contour. In Figure 6.8d, all the peaks between frames 120-155 are spurious peaks and they are discarded because their $sv(n)$ values are below the SVth of this word. These peaks occur along the geminated consonant /f/, which is regarded as one sub-word unit, according to the SV contour, with a duration of 40 frames. This latter problem is resolved during the segmentation process where the duration is consulted to decide whether a certain consonant is single or geminated. This is explained later, in Chapter 7.

Figure 6.9 shows graphs for the word 'walbuḥuuḥi', where a) shows the speech signal, b) the V-UV-S contour, c) the ES contour, and d) the SV contour. In Figure 6.9d, all the spurious peaks which occur between the peak pairs P2-P3, P6-P7, and P8-P9 are ignored because they all have $sv(n)$ values below the SVth for this word. Unfortunately, the peak at frame 40 which is associated with the transition between the phonemes /a/ and /l/ is also discarded for the same reason. Nevertheless, the boundary between /a/ and /l/ is predicted later on in the error correction process described in Chapter 7, by employing durational information from the ES and SV contours.

Figure 6.10 shows graphs for the word 'maktabatan', where a) shows the speech signal, b) the V-UV-S contour, c) the ES contour, and d) the SV contour. The SV contour of this word shows some spurious peaks before frame number 20 and after frame number 212. Those peaks are neglected because they are in the silence regions outside the word boundaries according to the V-UV-S contour. As a result, all peaks outside the word endpoints are neglected regardless of their $sv(n)$ values. It should be mentioned here that the above-mentioned peaks in Figure 6.10d all have $sv(n)$ values below the SVth of their word.

Figure 6.11 shows graphs for the word 'jaḥsub', where a) shows the speech signal, b) the V-UV-S contour, c) the ES contour, and d) the SV contour. The SV contour displays peak (P4) at frame number 62 which is related to the transition between the two unvoiced phonemes /ḥ/ and /s/. Thus, the SV contour shows 7 segments associated with the six phonemes in this word, where the last two segments are associated with the plosive phoneme /b/.

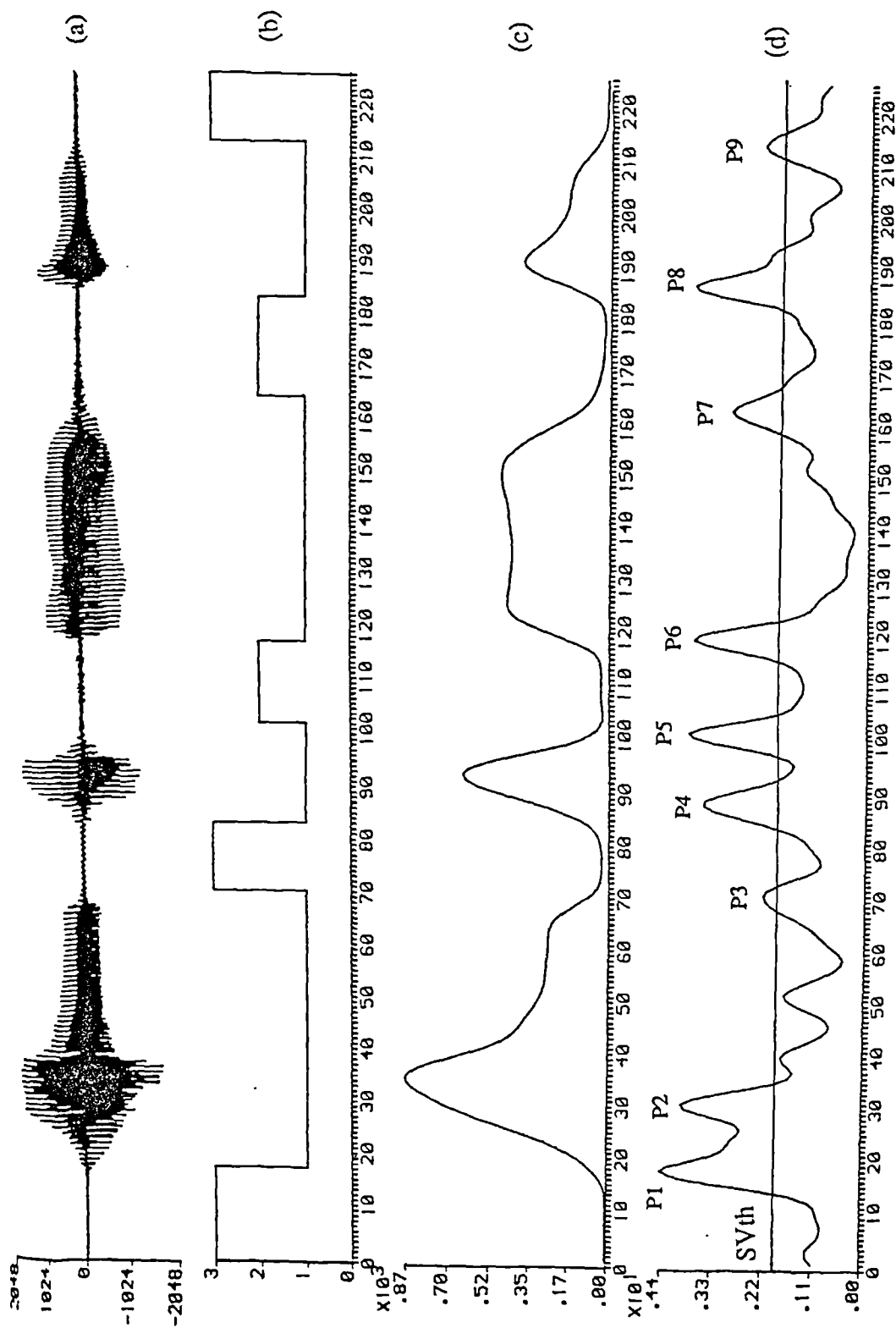


Fig. 6.9 Graphs for the word 'walbuhuuθi'
 a) the speech signal, b) the V-UV-S contour, c) the ES contour, d) the SV contour

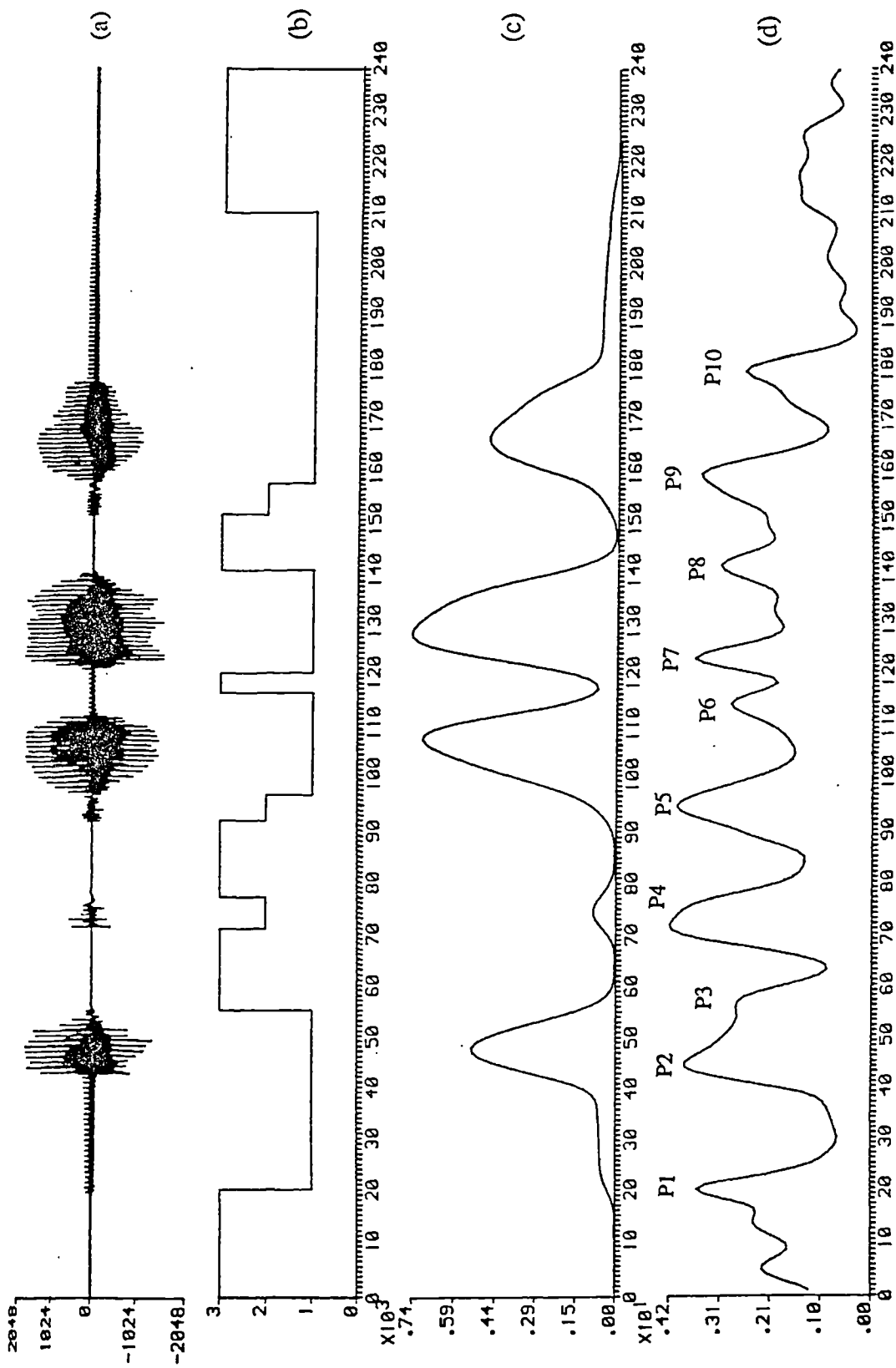


Fig. 6.10 Graphs for the word 'maktabatan'

a) the speech signal, b) the V-UV-S contour, c) the ES contour, d) the SV contour

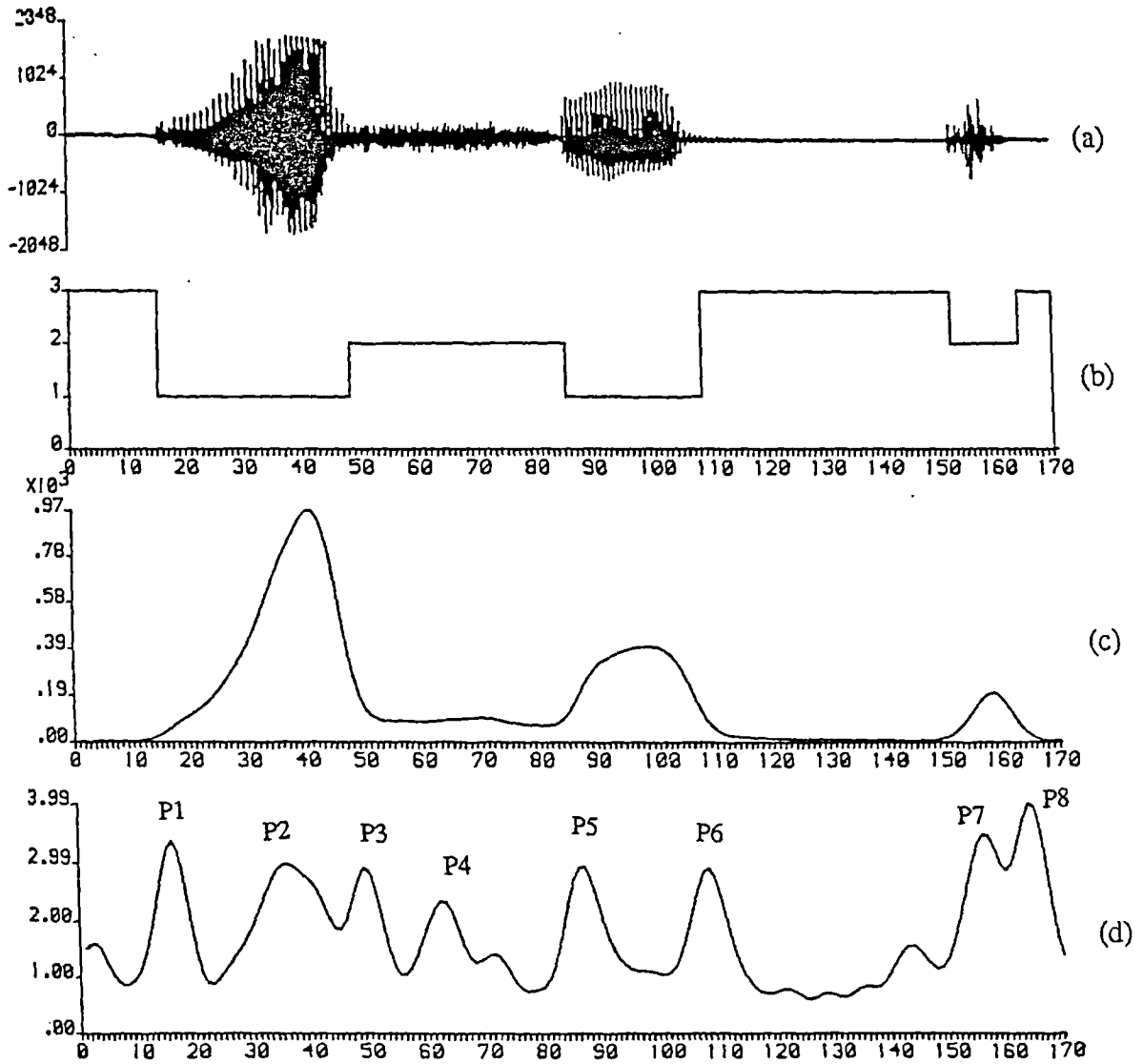


Fig. 6.11 Graphs for the word 'jahsub'
 a) the speech signal, b) the V-UV-S contour
 c) the ES contour, d) the SV contour

Figure 6.12 shows graphs for the word 'laakinna', where a) shows the speech signal, b) the V-UV-S contour, c) the ES contour, and d) the SV contour. This word contains a geminated voiced consonant /n/. The SV contour of this word does not show any peak along the duration of this consonant. This problem is also resolved by employing durational information in the error correction process described in the next chapter.

6.5 Summary

The concept of the spectral variation function is introduced in this chapter. This function is computed by using the mel-frequency cepstral parameters. The extraction of these parameters from the speech signal is also demonstrated. Six mel-frequency cepstral parameters are used in the computation of the spectral variation function along each word. It has been shown that the spectral variation contour of a certain word displays peaks corresponding to the transitional regions between adjacent sub-word units in that word. These units are mainly single phonemes, but they may comprise more than one phoneme.

The peak points of the SV contour of a certain word determine the boundaries of the segments contained in that word. Thus, the SV contour provides a tool for an automatic segmentation algorithm. In this algorithm the segments of a certain word are determined regardless of their content or identity.

The SV contour may sometimes fail to locate the boundaries between adjacent phonemes due to:

- non-existence of a clear or prominent transition in the spectrum
- the duration of a phoneme being too small for the computation of the $sv(n)$ function and performing the smoothing process.

The problems of missing boundaries and/or the presence of extra peaks which refer to false boundaries, are tackled in the segmentation and error correction procedure presented in Chapter 7, where durational information and phonological constraints are employed.

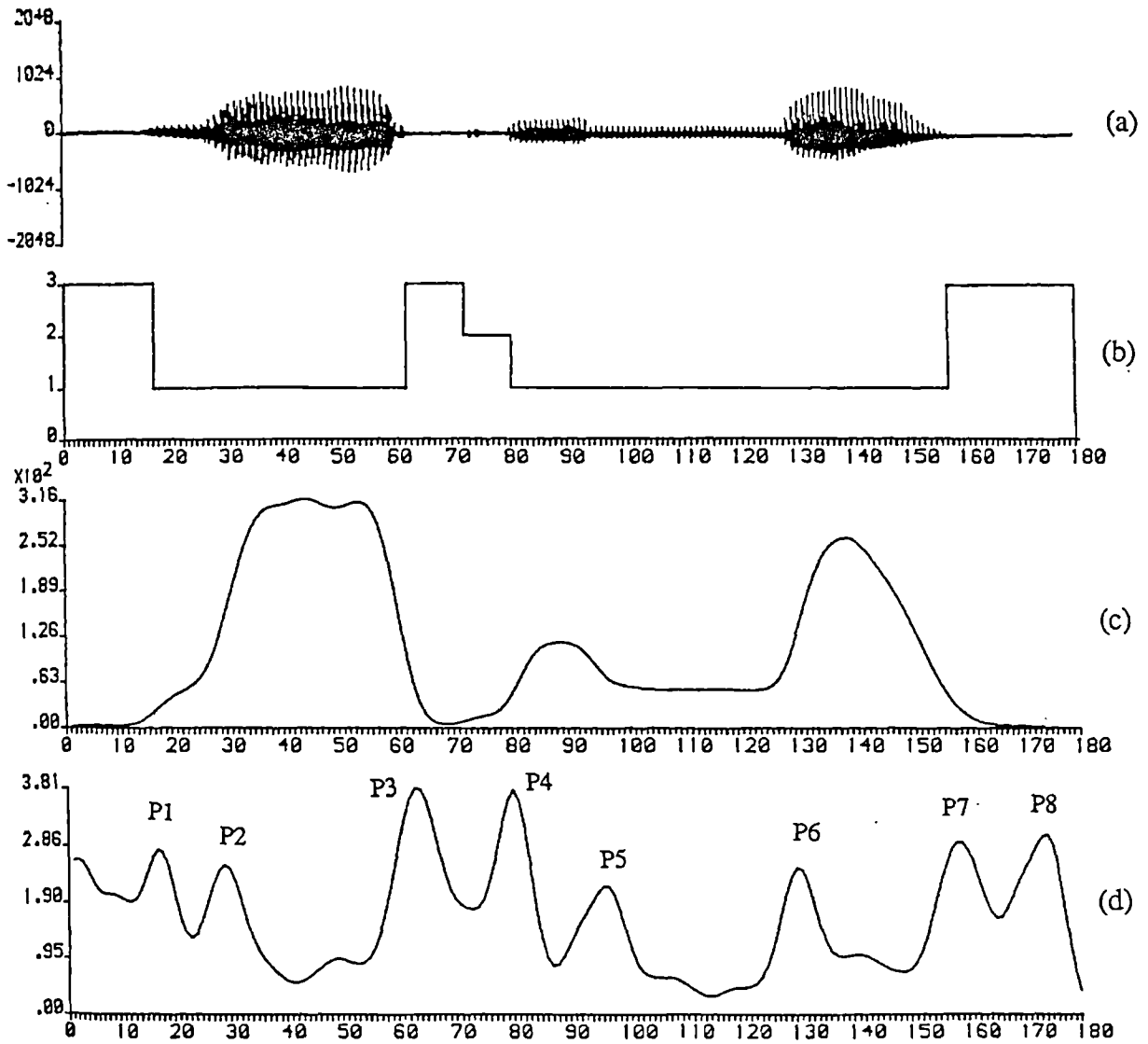


Fig. 6.12 Graphs for the word 'laakinna'
 a) the speech signal, b) the V-UV-S contour
 c) the ES contour, d) the SV contour

Chapter 7**Segmentation and Error Correction****7.1 Introduction**

Segmenting and labelling the speech signal in terms of broad phonetic classes (for consonants) according to the classification scheme number 7 (see Section 4.3.1), are described in this chapter. In this scheme, the consonants are classified into four categories, namely:

- Voiced plosive 'VP'
- Unvoiced plosive 'UP'
- Unvoiced fricative 'UF'
- Voiced consonant 'VC'

The vowels are also classified into six vowels (three short and three long), and they are labelled according to their phonetic symbol as follows:

'a' , 'u' , 'i' , 'aa' , 'uu' , and 'ii'

This leads to a set of 10 different labels. Thus, each word is to be described by a string of labels selected from this set.

Up to this stage, the following steps are achieved:

- V-UV-S segmentation
- Vowel recognition
 - Vowel representative frame (VRF)
 - Vowel estimated duration (VED)
 - Vowel identity (vowel label)
- The SV contour which gives the boundaries between adjacent sub-word units (mainly phonemes)

It can be seen that most of the information which is necessary to proceed with the segmentation and labelling processes is available.

Before carrying on with the labelling process, plosive detection is performed by using both the result of the V-UV-S segmentation and the pitch detection algorithm. In this respect, the silence and unvoiced segments associated with plosive phonemes are detected and given special codes to be used later on in the labelling process.

The initial segmentation results are passed through a special error correcting procedure in order to finalise these results. In this procedure, durational information in addition to some phonological constraints are used to reach an acceptable or legitimate syllabic structure as explained in the coming sections.

7.2 Plosive Detection

As shown in Chapter 3, the spectrogram of the voiced plosive phoneme (or voiced stop) is characterised by a voice bar along the base line followed by a sudden burst noise. An unvoiced plosive phoneme is characterised by a silence gap followed by a sudden burst noise (see Figure 3.4). In the V-UV-S segmentation procedure (see Section 5.2), the speech signal is highpass-filtered to 300 Hz which is above the limit of the voice bar associated with voiced plosive sounds. All plosive sounds are then shown on the V-UV-S contour by a silence segment followed generally by a short unvoiced segment. This latter segment represents the burst noise associated with plosive sounds.

The first step in the plosive detection algorithm is to analyse the region immediately following each silence segment on the V-UV-S contour, to determine the presence and location of a burst. Then, an initial decision is made, based on duration, as to whether an unvoiced segment following the silence is a fricative, or simply the burst (or aspiration) following the silence gap. The next step is to determine whether the silence gap belongs to a voiced or an unvoiced plosive. This is achieved by checking the presence of voicing at a low frequency using the results of the pitch detection algorithm.

a) Burst Detection

Burst detection is accomplished by analysing the energy of a few frames near the end of the silence segment. The speech signal is first pre-emphasised (see Eq. (5.9)) by

a first-order filter, then the energy is computed for non-overlapping frames of 64 samples each (the overlapping is avoided in order not to obliterate the burst). Then, the second derivative of the energy function is calculated and called E_b . The values of E_b for the first and the second unvoiced frames following the silence are checked. If either of them exceeds a certain burst threshold B_{th} , the location of the relevant frame is selected as the burst pointer. B_{th} is chosen empirically equal to 500. If the unvoiced segment following the silence exceeds 12 frames or 76 msec (16 frames for silence segment at word-final position), the segment is considered as a fricative segment. The entire sequence of silence and burst represents plosive sound.

Burst detection is carried out on all unvoiced segments following silence segments including the silence at word-initial position, where a word may start with a plosive phoneme. Sometimes the silence segment related to a plosive phoneme is not followed by an unvoiced segment related to the plosive's burst. This is actually the case of an unreleased burst (not pronounced) which occurs especially at word-final position. This will not create a problem for plosive detection when a plosive occurs at word-medial position, unlike the cases at word-initial or word-final positions. Such words which have a plosive at word-final position can be represented by two entries in the lexicon (i.e., with and without the final plosive phoneme).

b) Voiced Plosive Detection

Silence segments along a word are tested to determine the presence of voicing at low frequencies below 300 Hz. The presence of voicing is determined by checking two already available measures, namely the energy E in the range 60-4800 Hz (see Eq. (5.2) in Section 5.2.3), and the value of the pitch peak PPK which is measured in the pitch detection algorithm (see Section 5.2.2). Any silence frames whose E values exceed twice the energy of the background noise, and whose PPK exceed 0.45, are relabelled as voiced frames. If the number of frames (within a silence segment) satisfying this condition exceeds a certain threshold VP_{th} , the whole silence segment is then considered as voiced segment. VP_{th} is chosen equal to :

- 4 frames in the case of having a silence segment surrounded by two voiced segments (i.e., plosive phoneme surrounded by two vowels).
- 6 frames in the case of having a segment (or plosive phoneme) at word-initial or word-medial position.

- 10 frames in the case of having a silence segment at word-final position (where the burst of the plosive phoneme might be unreleased).

c) Results

In the V-UV-S segmentation process, silence, unvoiced, and voiced segments are given the codes '3', '2', and '1' respectively. The silence and unvoiced segments associated with plosive sounds are recoded as follows:

- '4' for silence segment associated with unvoiced plosive phoneme.
- '5' for silence segment associated with voiced plosive phoneme.
- '6' for unvoiced segment related to the burst associated with plosive phoneme.

Finally, any short unvoiced segment (less than 5 frames) at word-initial position (at the beginning of a word) followed by a voiced segment is eliminated if it does not pass the burst test. This segment is relabelled as a voiced segment and added to the following voiced segment. Such segments might have been generated by breathing noise.

7.3 Segmentation and Labelling

Figure 7.1 shows a block diagram of the processes involved in the segmentation and labelling procedure.

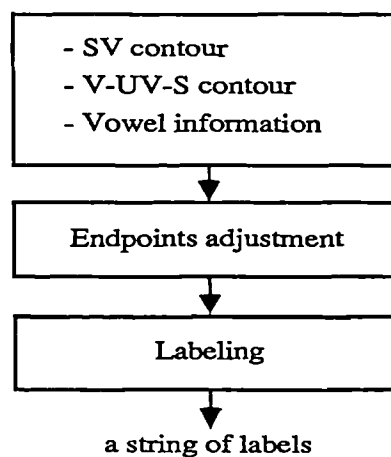


Fig. 7.1 Segmentation and labelling procedure

The first block in this diagram contains all the initial processes necessary for the segmentation. The results of these processes are supplied to the labelling algorithm. These results contain the VRFs, VEDs, vowel identities, the boundaries between sub-word units, and the updated version of the V-UV-S segmentation results which include any available plosive information.

Before labelling, the endpoints of the SV contour and the V-UV-S contour of each word are aligned to each other. Then, the labelling process is carried out according to the chosen classification scheme. The initial labelling results are checked to verify their correctness by using durational and phonological information as explained in Section 7.4.

7.3.1 Endpoint Adjustment

In this stage, a special process is performed to align between the beginning and the end of both the V-UV-S contour and the SV contour. Figure 7.2 displays graphs for the word 'saaka', where a) shows the speech signal, b) the V-UV-S contour, c) the ES contour, and d) the SV contour. The SV contour of this word shows five peaks at the following frame numbers:

| | | | | |
|----|----|-----|-----|-----|
| P1 | P2 | P3 | P4 | P5 |
| 15 | 58 | 111 | 131 | 150 |

and the V-UV-S contour gives the following segments:

| between frames | label | code | updated code | |
|----------------|-------|------|--------------|-------------------------------------|
| 1 - 22 | S | 3 | 3 | |
| 23 - 58 | UV | 2 | 2 | |
| 59 - 108 | V | 1 | 1 | |
| 109 - 126 | S | 3 | 4 | silence segment of unvoiced plosive |
| 127 - 131 | UV | 2 | 6 | burst noise of unvoiced plosive |
| 132 - 155 | V | 1 | 1 | |
| 156 - 168 | S | 3 | 3 | |

According to the V-UV-S contour this word begins at frame 23, while the SV contour shows a peak at frame 15. The reason for this mismatch between the edges of the two contours is due to the presence of an unvoiced fricative /s/ at the word-initial position.

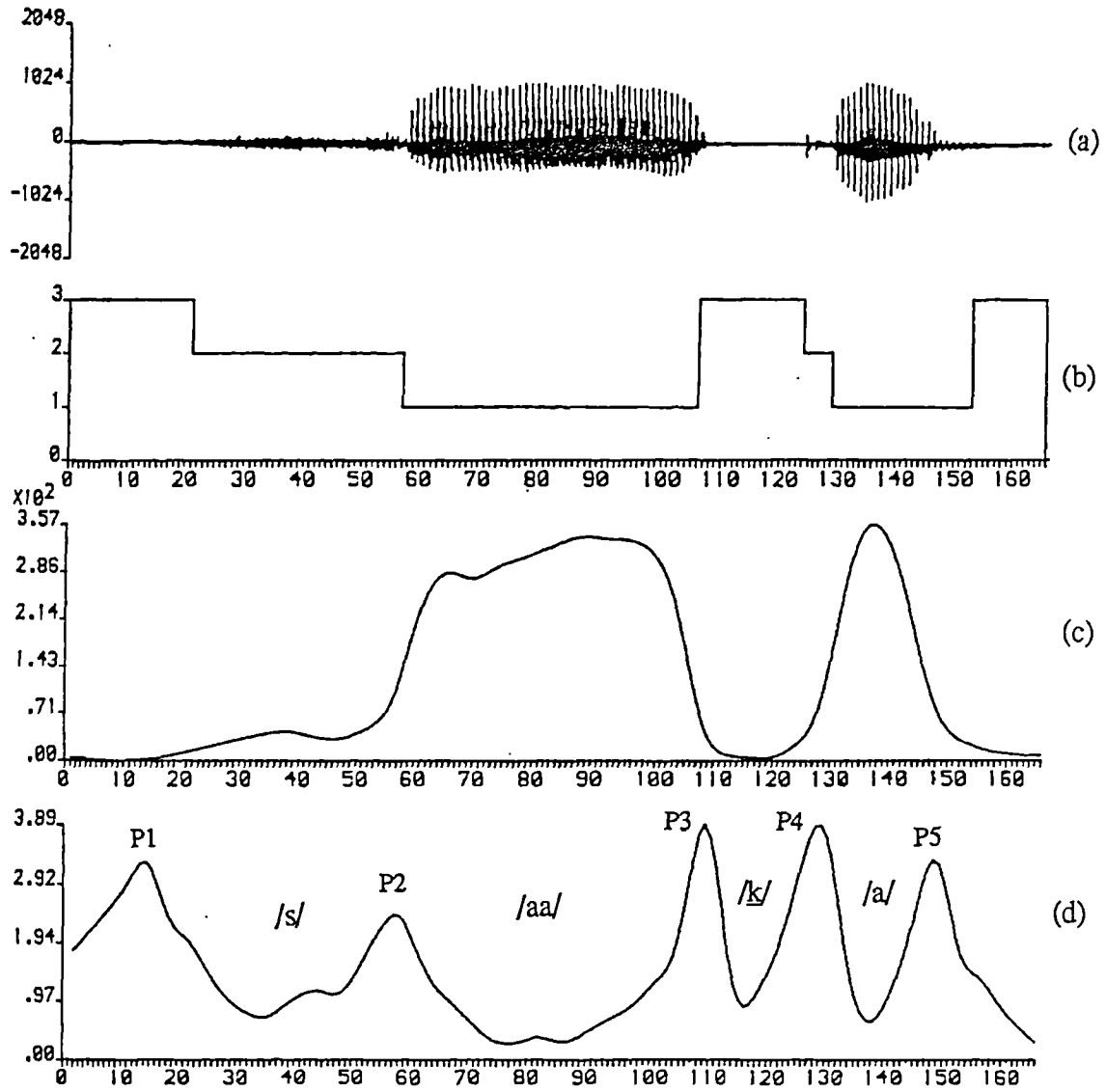


Fig. 7.2 Graphs for the word 'saaka'
 a) the speech signal, b) the V-UV-S contour
 c) the ES contour, d) the SV contour

Sometimes, it is difficult to detect the exact beginning point of a word if it begins with a weak fricative. The same problem arises at the end of this word, where the SV contour shows a peak at frame 150, while the last silence segment starts at frame 156. This latter case occurs due to the breathing noise which is present at the end of this word.

Nevertheless, general rules are applied to each word to determine the exact endpoints of that word. Two cases are distinguished as illustrated in Figure 7.3. In this figure both the cases (a) and (b) have Δ frames ($\Delta = |X(\text{IB}) - X(\text{P1})|$) between the first peak of the SV contour and the beginning point (IB) at the V-UV-S contour. When Δ is less than or equal to 4 frames, the beginning point of a word is taken at P1 for the case (a) and at (IB) for the case (b). When Δ is more than 4 frames, the beginning point is taken at (IB) on the V-UV-S contour for the case (a) and a new peak is added before the first peak on the SV contour, while for the case (b) the first peak on the SV contour (P1) is shifted to a new position which coincides with the point (IB) on the V-UV-S contour. The same rules are applied at the end of a word.

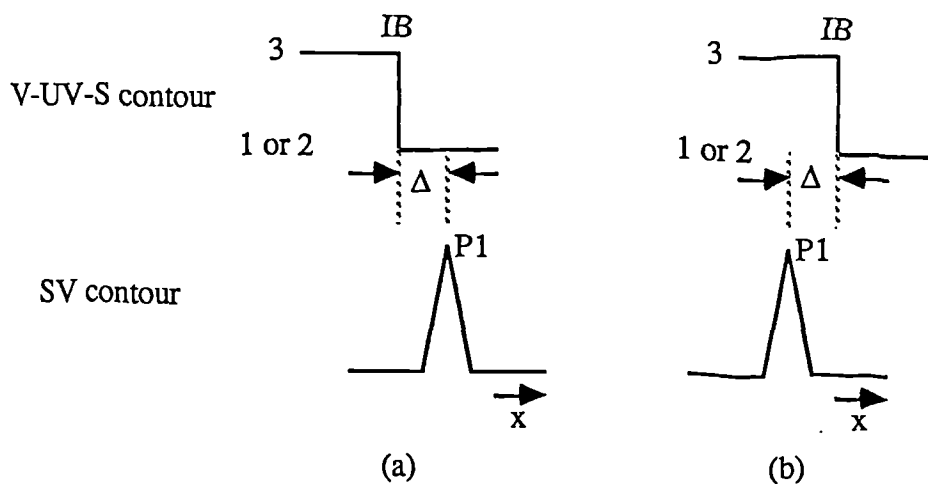


Fig. 7.3 Endpoint adjustment

For the word 'saaka', the beginning point is taken at frame 23, and the end point at frame 150, (see Figure 7.2). Figures 6.6 and 6.7 in Chapter 6 display similar cases, where the starting point of each word is taken at the edge of the first silence segment, and a new peak is created at this point on the SV contour.

In general, differences (Δ s) between the peak points of the SV contour and their counterparts the transitional points on the V-UV-S contour (i.e., the boundaries of the segments on the V-UV-S contour) may occur due to the following reasons:

- smoothing both the energy outputs of the filter bank and the mel-frequency cepstral parameters
- inaccuracy in the V-UV-S decision or due to some error which might occur as a result of the non-linear smoothing applied to the V-UV-S contour
- the inherent limitation of the defined SV function and its computation over a window of $2L+1$ frames

However, differences up to 3 frames are tolerated where L is taken equal to 2 frames and the smoothing is carried out mainly over 3 frames.

7.3.2 Labelling

The labelling process is explained by the demonstration of several examples. A special table is created for each word and called the segmentation result table. This table contains the previously obtained results from the vowel recognition procedure, the V-UV-S segmentation procedure, the plosive detection process, and the segments of the SV contour. The results of the labelling process are also added to this table.

Table 7.1 shows the segmentation results of the word 'saaka'. The results are given graphically in Figure 7.2. This table has 5 main columns. The first one contains the serial number of the segments along the SV contour, while the second main column contains the boundaries (i.e., the peak points of the SV contour given by their frame numbers) and the length of those segments (given by the number of frames between the two peaks of each segment). The third main column has four sub-columns which contains the results of the vowel recognition procedure. The first sub-column (which is called Cd1) displays the code '1' if a VRF lies within the boundaries of that segment, and such segment is called a vowel segment. The other three sub-columns display the location of the vowel representative frame, the vowel estimated duration (in frames), and the vowel identity if the current segment is a vowel segment. It is assumed that each segment along the SV contour may associate with one or two segments along the V-UV-S contour, therefore the fourth main column contains two sub-columns (which are called Cd2 and Cd3) to indicate the V-UV-S decision. The fifth main column

displays the labelling result based on the information given in the preceding three main columns.

| (1) seg. no. | (2) segment boundaries | | | (3) vowels | | | | (4) updated V-UV-S | | (5) label |
|--------------------|------------------------------|-----|--------|---------------|-----|-----|------|--------------------------|-----|--------------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 23 | 57 | 35 | 0 | - | - | - | 2 | 2 | UF |
| 2 | 58 | 110 | 43 | 1 | 81 | 44 | /aa/ | 1 | 1 | aa |
| 3 | 111 | 130 | 20 | 0 | - | - | - | 4 | 6 | UP |
| 4 | 131 | 150 | 20 | 1 | 141 | 12 | /a/ | 1 | 1 | a |

Table 7.1 Segmentation results for the word 'saaka'

In Table 7.1, the peak points of the SV contour of Figure 7.2d and the duration (in frames) between two successive peaks are filled in the second main column. The third column is filled with the vowel representative frames which are detected from the ES contour of Figure 7.2c, the estimated duration of each vowel, and the vowel identities. The fourth column is filled with codes which are taken from the V-UV-S contour of Figure 7.2b along with the results of the plosive detection process.

The first segment in Table 7.1 is given the code '2' in both Cd2 and Cd3 columns (the fourth main column) meaning that this segment is completely an unvoiced segment, where the segment's boundaries on the SV contour match its boundaries on the V-UV-S contour. As we said earlier any differences (Δ) between the location of the peak points on the SV contour and their counterparts the transitional points (between two segments) along the V-UV-S contour, of less than or equal to 4 frames are tolerated. Thus, This segment is assigned the label 'UF'.

The second segment in this table is given the code '1' for both Cd2 and Cd3 columns (voiced segment) and considered as a vowel segment (column Cd1 shows the code '1' for this segment), because the vowel representative frame lies between its boundaries, and is almost at the centre of this segment (at frame 81). Thus, this segment is assigned

the label 'aa' which is equivalent to the vowel identity given in column (ID) by the vowel recognition procedure.

The third segment in the table is given the code '4' in the Cd2 column and the code '6' in the Cd3 column. This means that the silence in the first part is related to the gap associated with the unvoiced plosive phoneme /k/, while the unvoiced segment in the second part is related to the burst noise associated with the plosive phoneme /k/. Thus, this segment is assigned the label 'UP'. The SV contour of this word (Figure 7.2d) did not display any peaks between the silence and unvoiced segments associated with the plosive phoneme /k/.

The fourth segment in the table is given the code '1' in both Cd2 and Cd3 columns and the Cd1 column has the code '1' (i.e., vowel segment). Thus, this segment is assigned the label /a/ which is equivalent to the vowel identity given in column (ID) by the vowel recognition procedure.

As a result, the word 'saaka' is described by the following string of labels:

UF-aa-UP-a its syllabic pattern /CVV-CV/

This syllabic pattern contains a legitimate syllabic structure according to the phonological rules given in Section 3.7.2. Nevertheless, the above string of labels is considered as an initial result. The final string of labels describing a certain word results after passing the initial results through a special correction procedure as will be explained in Section 7.4.

Another example is the segmentation results of the word 'maṭaaliba'. Table 7.2 illustrates the segmentation results of this word, and these results are given graphically in Figure 7.4. In the same way as in the previous example, Table 7.2 is filled with the peak points of the SV contour of Figure 7.4d, the results of the vowel recognition procedure (extracted from the ES contour of Figure 7.4c), and the updated V-UV-S segmentation results (which include the results of the plosive detection). Actually, filling the results in the fourth main column in the segmentation result table is the most difficult and critical process of the segmentation and labelling procedure. In this case, errors might occur due to the misalignment between the SV contour's peaks and their counterparts, the transitional points on the V-UV-S contour. This can occur especially at the transition between an unvoiced segment related to a plosive phoneme and the following voiced

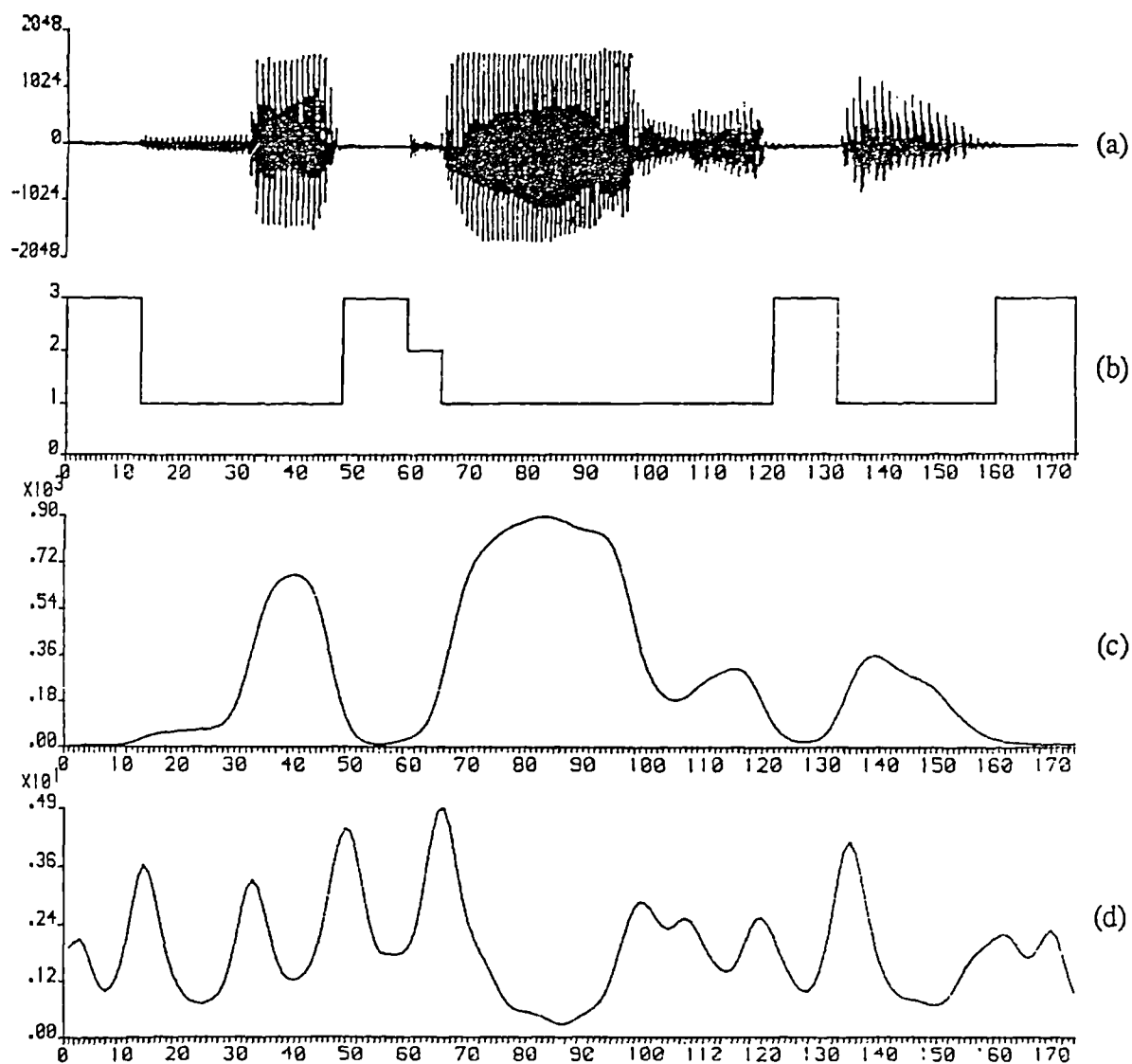


Fig. 7.4 Graphs for the word 'mataaliba'
 a) the speech signal, b) the V-UV-S contour
 c) the ES contour, d) the SV contour

segment related to a vowel or other voiced phoneme. This can be resolved by tolerating this kind of jitter. Column 5 of Table 7.2 gives the following string of labels for the word 'maṭaaliba':

VC-a-UP-aa-VC-i-VP-a its syllabic pattern /CV-CVV-CV-CV/

and the resultant syllabic pattern is legitimate according to the rules given in Section 3.7.2.

| seg. no. | segment boundaries | | | vowels | | | | updated V-UV-S | | label |
|----------|--------------------|-----|--------|--------|-----|-----|------|----------------|-----|-------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 15 | 33 | 19 | 0 | - | - | - | 1 | 1 | VC |
| 2 | 34 | 49 | 16 | 1 | 41 | 13 | /a/ | 1 | 1 | a |
| 3 | 50 | 66 | 17 | 0 | - | - | - | 4 | 6 | UP |
| 4 | 67 | 100 | 34 | 1 | 83 | 29 | /aa/ | 1 | 1 | aa |
| 5 | 101 | 108 | 8 | 0 | - | - | - | 1 | 1 | VC |
| 6 | 109 | 121 | 13 | 1 | 117 | 13 | /i/ | 1 | 1 | i |
| 7 | 122 | 136 | 15 | 0 | - | - | - | 5 | 5 | VP |
| 8 | 137 | 162 | 26 | 1 | 145 | 14 | /a/ | 1 | 1 | a |

Table 7.2 Segmentation results for the word 'maṭaaliba'

The initial labelling algorithm is based on the codes given in the third and fourth main columns of the segmentation result table of a certain word. Figure 7.5 illustrates a flow chart of the labelling algorithm. In this algorithm, the codes Cd1, Cd2, Cd3, and the vowel identity of each segment in the segmentation result table are tested to determine the label of that segment. So, if Cd1 is greater than zero, this means that the current segment is a vowel segment and it is labelled according to the vowel identity given in column (ID) in the result table. But if Cd1 is equal to zero, then both codes Cd2 and Cd3 are checked. If both codes are '1' the segment is assigned the label 'VC' (voiced consonant), while if both codes are '2' the segment is assigned the label 'UF' (unvoiced fricative). When Cd2 is '4' the segment is tagged as 'UP' (unvoiced plosive phoneme), or if it is

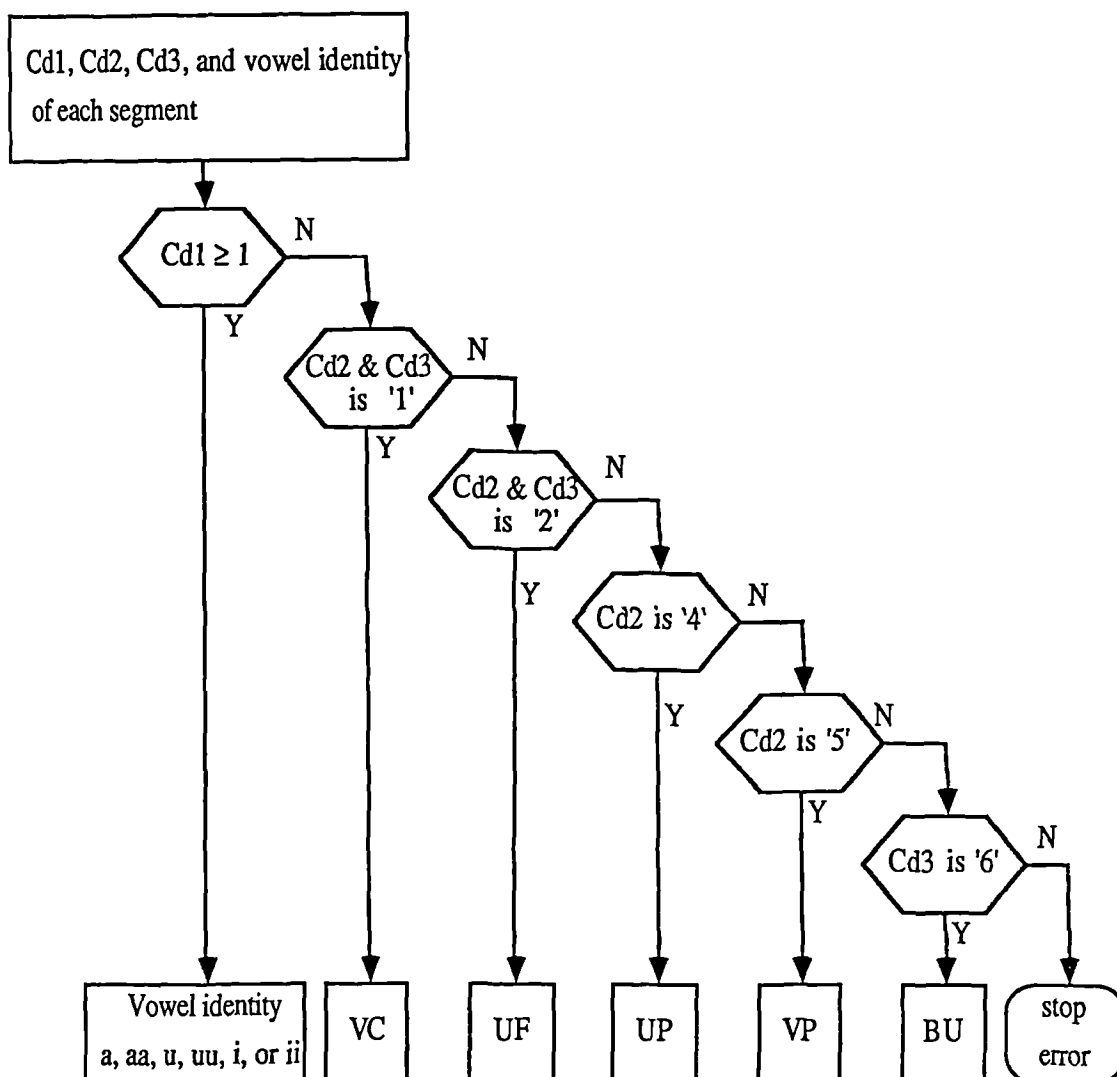


Fig. 7.5 Flow chart of the labelling algorithm

'5' the segment is tagged as 'VP' (voiced plosive). When Cd2 is neither '4' nor '5' then Cd3 is checked: if it is '6' the segment is labelled as burst segment 'BU' otherwise the algorithm shows a state of error. The last case occurs when the SV contour displays a peak between the silence and the unvoiced segments associated with a plosive phoneme.

The previous two examples have shown almost perfect segmentation and labelling results. This is not always the case, since some errors may occur due to the following reasons:

- the presence of a geminated consonant, and the SV contour shows one segment for this consonant
- missing a glottal phoneme at word-initial position.
- the SV contour shows two segments for a plosive phoneme
- missing voiced consonant segments
- the SV contour displays two segments relating to one vowel

The following paragraphs demonstrate such types of errors that arise during the segmentation procedure, and highlight some parameters which are useful for correcting these errors.

The first example is the segmentation of the word 'ʔaddiraasaati'. Table 7.3 shows the segmentation results for this word, and these results are given graphically in Figure 7.6. Table 7.3 shows that this word has 9 segments according to the SV contour of Figure 7.6d. This contour displays a peak at frame 191 related to the transition between silence and unvoiced parts of the unvoiced plosive phoneme /t/, and hence this phoneme is represented by two segments in the result table. These two segments are combined together into one segment, as will be explained, in the plosive correction algorithm (Section 7.4.2).

| seg. no. | segment boundaries | | | vowels | | | | updated V-UV-S | | label |
|----------|--------------------|-----|--------|--------|-----|-----|------|----------------|-----|-------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 13 | 25 | 13 | 1 | 19 | 8 | /a/ | 1 | 1 | a |
| 2 | 26 | 60 | 35 | 0 | - | - | - | 5 | 6 | VP |
| 3 | 61 | 78 | 18 | 1 | 72 | 16 | /i/ | 1 | 1 | i |
| 4 | 79 | 117 | 39 | 1 | 95 | 32 | /aa/ | 1 | 1 | aa |
| 5 | 118 | 137 | 20 | 0 | - | - | - | 2 | 2 | UF |
| 6 | 138 | 179 | 42 | 1 | 156 | 29 | /aa/ | 1 | 1 | aa |
| 7 | 180 | 191 | 12 | 0 | - | - | - | 4 | 4 | UP |
| 8 | 192 | 200 | 9 | 0 | - | - | - | 6 | 6 | BU |
| 9 | 201 | 225 | 25 | 1 | 210 | 20 | /i/ | 1 | 1 | i |

Table 7.3 segmentation results for the word 'ʔaddiraasaati'

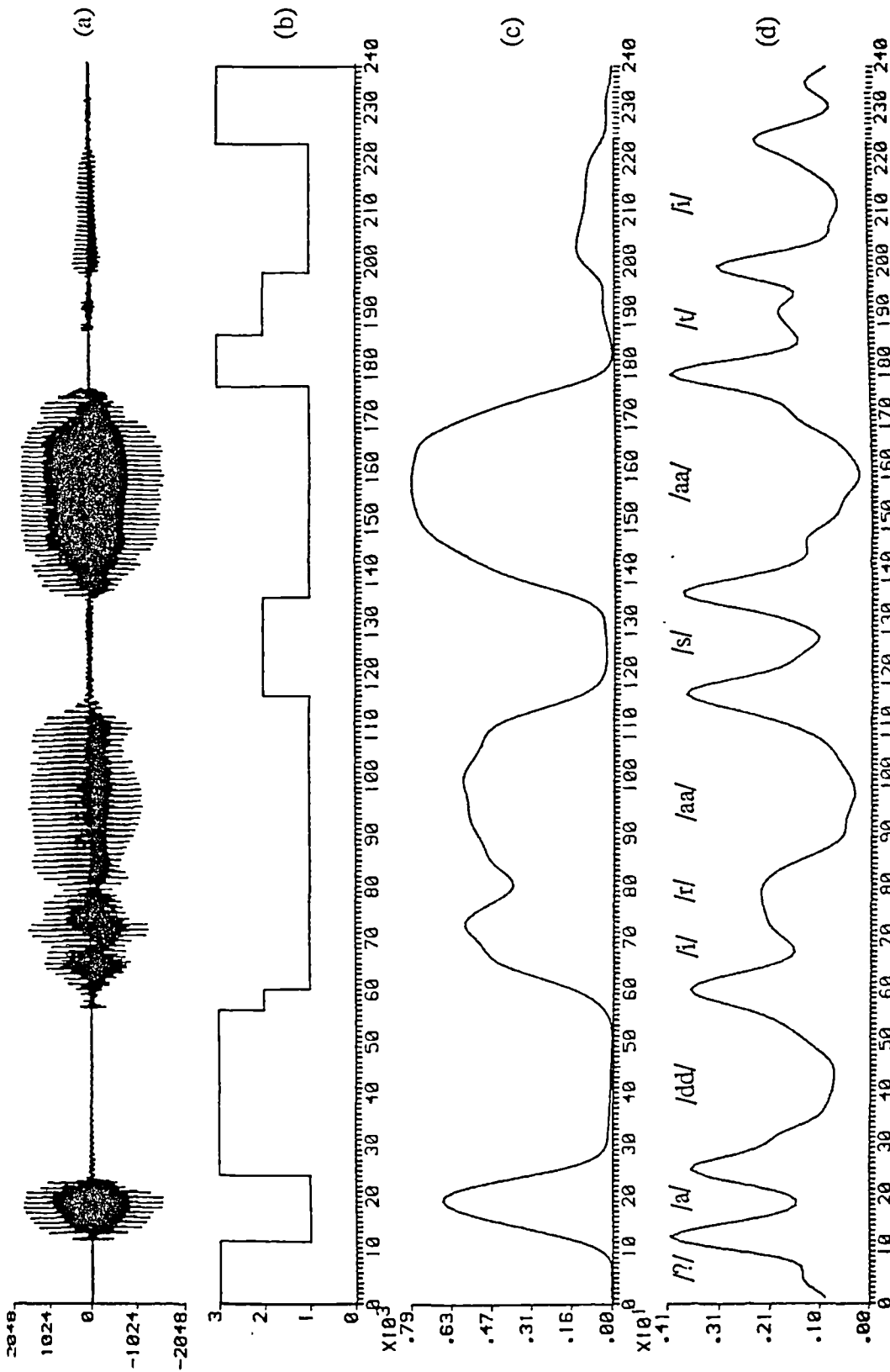


Fig. 7.6 Graphs for the word 'addiraasaati'

a) the speech signal, b) the V-UV-S contour, c) the ES contour, d) the SV contour

However, this word is described by the following string of labels (after removing the burst segment):

a-VP-i-aa-UF-aa-UP-i its syllabic pattern /N-CV-VV-CVV-CV/

The actual phonetic description of this word is as follows:

UP-a-VP-VP-i-VC-aa-UF-aa-UP-i its syllabic pattern /CVC-CV-CVV-CVV-CV/

By comparing the actual string of labels with the result of the labelling process we notice that 3 labels are missing. Those labels and their related phonemes are as follows:

- 'UP' related to unvoiced plosive or glottal phoneme /ʔ/ which is the leading consonant of the first syllable.
- 'VP' related to the voiced plosive phoneme /d/, which is the trailing consonant of the first syllable.
- 'VC' related to the voiced consonant /r/, which is the leading consonant of the third syllable.

These results arise because the SV contour of Figure 7.6d failed to display peaks (or transitions) related to the above missing three segments.

Actually, those missing labels (or segments) can be recovered by the correction procedure which uses durational information given in the result table. The segmentation errors of the word 'ʔaddiraasaati' are described in the following paragraphs. Also some parameters which are useful in the correction procedure are highlighted.

The duration or the length (in frames) of each segment is given in the column 'length' of the result table. By looking at the duration of the consonantal segments (i.e., non-vowel segments) in Table 7.3, we notice the following. The second segment which is labelled as 'VP' has a duration of 35 frames (224 msec), while segments 7 and 8 which are labelled as 'UP' & 'BU' both have a duration of less than 20 frames. Also, none of the other consonants in this word has a duration of more than 20 frames. This leads to the assumption that the 'VP' consonant is a geminated consonant (see Section 3.7.3). Thus, the second segment is split into two segments each of which has the same 'VP' label, and the modified string of labels becomes as follows:

a-VP-VP-i-aa-UF-aa-UP-i its syllabic pattern /VC-CV-VV-CVV-CV/

The first phoneme of this word is the glottal phoneme /ʔ/. This phoneme is described as an unvoiced plosive phoneme. The phoneme occurs mainly in word-initial position, and often it is not stressed during pronunciation. Thus, the burst relating to plosive phonemes often does not occur at the beginning of the V-UV-S contour of such words that start with a glottal phoneme. As explained in Section 3.7.1, any syllable in the Arabic language must start with a single consonant, therefore whenever we have a string of labels for a certain word which starts with a vowel, an unvoiced plosive label should be added before the first vowel label. For the word under discussion, Figure 7.6a shows a very short burst segment at the beginning of this word, but the V-UV-S detection algorithm failed to detect this burst. Thus, an unvoiced plosive label 'UP' is added before the first vowel. The string of labels describing this word becomes:

UP-a-VP-VP-i-aa-UF-aa-UP-i its syllabic pattern /CVC-CV-VV-CVV-CV/

It can be seen that the third syllable of this word has no initial consonant, and this results in an unacceptable syllabic structure. This has led to the idea of correcting this structure by adding a voiced consonant label 'VC' between the two vowel /i/ and /aa/. Then, the string of labels of this word becomes:

UP-a-VP-VP-i-VC-aa-UF-aa-UP-i its syllabic pattern /CVC-CV-CVV-CVV-CV/

It can be seen from the SV contour of Figure 7.6d that the bell shape at frame 79 is wider (more flat) than other shapes at other peaks, because this shape combines the two peaks, one at frame 79 and the other at frame 72. This latter peak has been smeared out by the smoothing processes which are applied to the system's parameters. Note that the consonant /r/ has a very short duration in this word, as is usually the case when it occurs in intervocalic location.

In the above example, it is shown that three correction processes should be carried out on the original segmentation results in order to reach the final acceptable (legitimate) syllabic structure. Thus, so far three correction processes must be designed, namely geminated consonant correction, plosive correction and correction according to the syllabic pattern. Another necessary correction process is demonstrated in the following example.

Table 7.4 shows the segmentation results for the word 'jafæaluuna', and these results are given graphically in Figure 7.7. According to this table, this word has eight segments, so the labelling algorithm gives a string of eight labels as follows:

VC-a-UF-a-VC-uu-VC-a its syllabic pattern /CV-CV-CVV-CV/

| seg. no. | segment boundaries | | | vowels | | | | updated V-UV-S | | label |
|----------|--------------------|-----|--------|--------|-----|-----|------|----------------|-----|-------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 18 | 32 | 15 | 0 | - | - | - | 1 | 1 | VC |
| 2 | 33 | 44 | 12 | 1 | 36 | 11 | /a/ | 1 | 1 | a |
| 3 | 45 | 71 | 17 | 0 | - | - | - | 2 | 2 | UF |
| 4 | 72 | 100 | 29 | 1 | 93 | 16 | /a/ | 1 | 1 | a |
| 5 | 101 | 112 | 12 | 0 | - | - | - | 1 | 1 | VC |
| 6 | 113 | 153 | 41 | 1 | 127 | 43 | /uu/ | 1 | 1 | uu |
| 7 | 154 | 160 | 7 | 0 | - | - | - | 1 | 1 | VC |
| 8 | 161 | 182 | 22 | 1 | 170 | 15 | /a/ | 1 | 1 | a |

Table 7.4 Segmentation results for the word 'jafæaluuna'

The resultant syllabic pattern appears to be a legitimate structure, but this does not mean that the segmentation result is correct, (where the actual syllabic pattern is /CVC-CV-CVV-CV/). From Table 7.4, we notice that the lengths of all the consonantal segments are within the normal limit (i.e., less than 20 frames in this word). For the vowel segments, we can write the following information:

| | | | | |
|--------------|-----|-----|------|-----|
| vowel | /a/ | /a/ | /uu/ | /a/ |
| length | 12 | 29 | 41 | 22 |
| VED | 11 | 16 | 43 | 15 |
| VED - length | -1 | -13 | 2 | -7 |

The last line above displays the difference between the defined vowel estimated duration from the ES contour (see Section 5.3.1a) and the segment length according to the SV

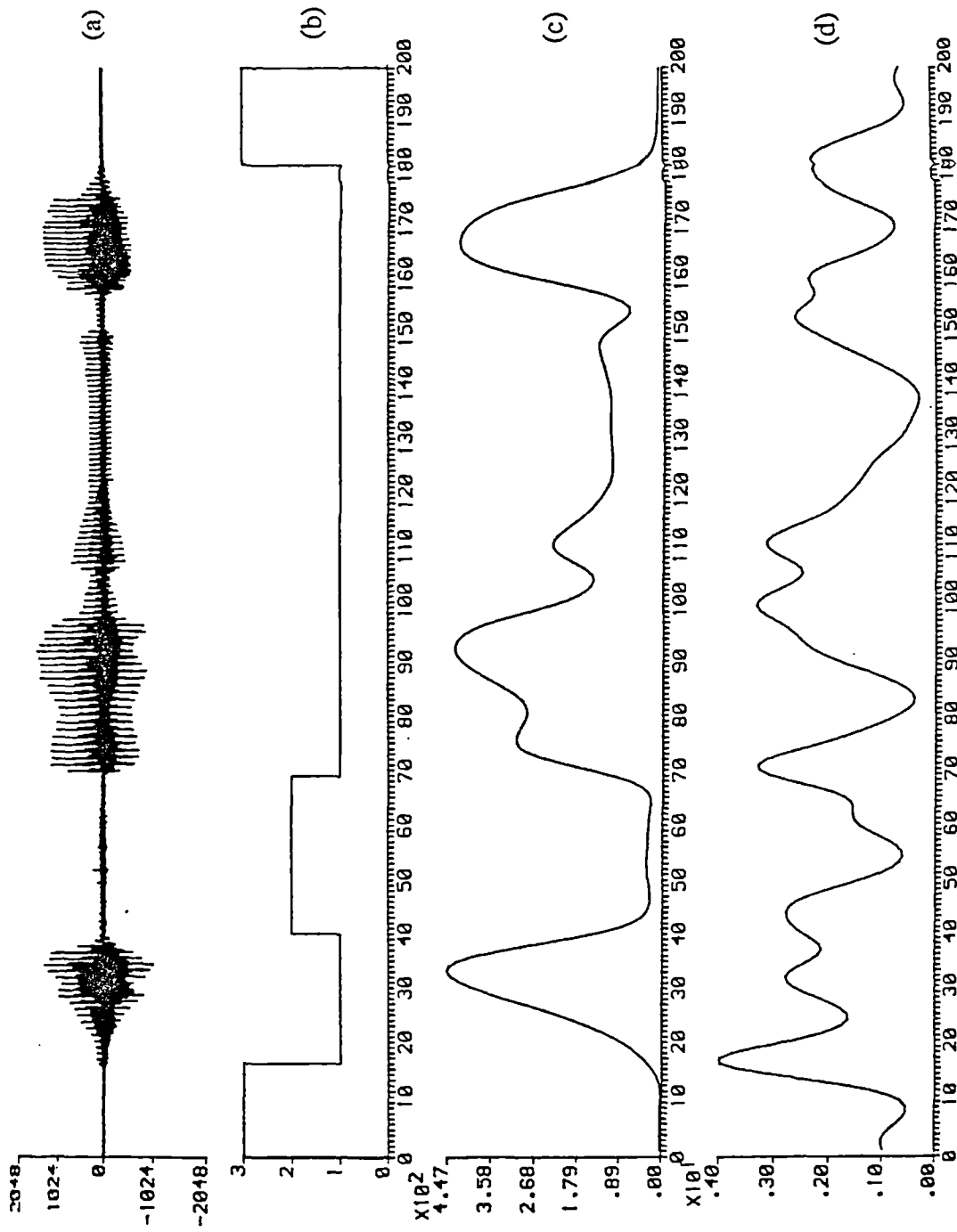


Fig. 7.7 Graphs for the word 'jafæluuna'
 a) the speech signal, b) the V-UV-S contour, c) the ES contour, d) the SV contour

contour. This difference is equal to -1, -13, 2, and -7 frames for the first, the second, the third, and the fourth vowel in this word, respectively. The difference of -13 frames related to the second vowel is actually below a certain limit (e.g., -7 frames). This leads to the assumption that the fourth segment contains a vowel and a voiced consonant. It can be clearly seen from Figure 7.7c and 7.7d, that the VRF (on the ES contour) of the second vowel lies in the right half of the fourth segment (on the SV contour). This leads to the conclusion that the assumed (missing) voiced consonant should precede this vowel. Then, the string of labels describing this word becomes:

VC-a-UF-VC-a-VC-uu-VC-a its syllabic pattern /CVC-CV-CVV-CV/

The difference (VED-length) for the third vowel is 2 frames, which means that the VED is longer than the related segment by just 2 frames. If this difference was above a certain limit (e.g., 5 frames for long vowel), the preceding (or the following, according to a certain rule) voiced consonant segment 'VC' would have to be added to the vowel segment. If we assume that this was the case for the long vowel /uu/ in the word 'jafæaluuna', this would have led to the syllabic pattern /CVC-CV-VV-CV/. The latter pattern is an illegitimate syllabic structure and could be corrected by adding a voiced consonant before the third vowel in this structure. Such a state is demonstrated in the following example.

Table 7.5 illustrates the segmentation results for the word 'tamsa^hhiina'. These results are given graphically in Figure 7.8.

According to this table, this word has 10 segments and is therefore described by a string of ten labels by the labelling algorithm as follows:

UP-a-VC-UF-a-UF-VC-ii-VC-a its syllabic pattern /CVC-CVC-CVV-CV/

The resultant syllabic pattern has an acceptable structure, but this does not mean that the segmentation result is correct, (where the actual syllabic pattern is /CVC-CV-CVV-CV/). It can be seen from Table 7.5 that the difference (VED-length = 45-35) for the third vowel /ii/ in this word is equal to 10 frames. This difference is above the allowed limit. This leads to the assumption that an extra peak (at frame 128) has been spotted on the SV contour along the vowel segment and has led to a false segment (the seventh segment in Table 7.5).

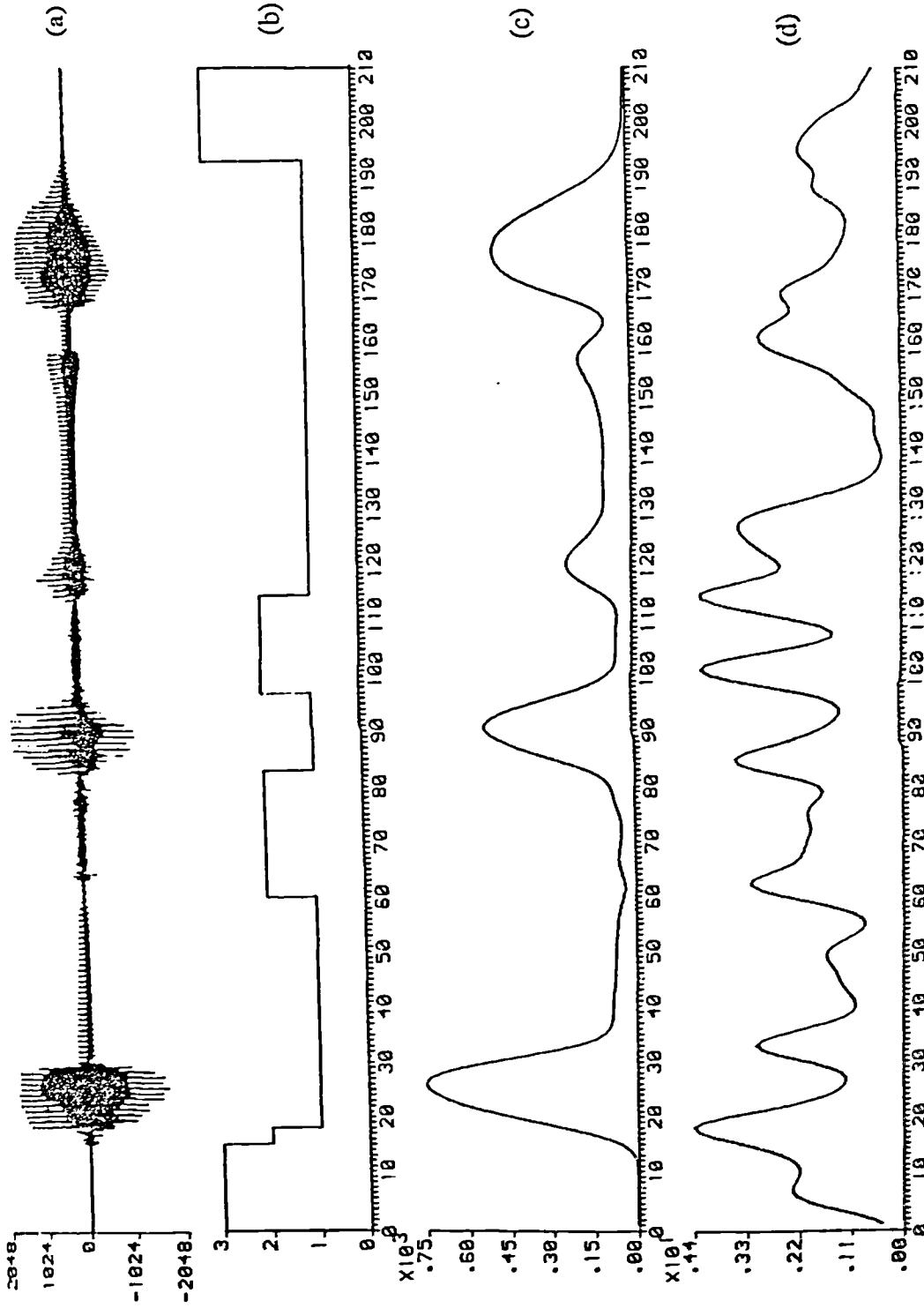


Fig. 7.8 Graphs for the word 'tamsahiina'
 a) the speech signal, b) the V-UV-S contour, c) the ES contour, d) the SV contour

| seg. no. | segment boundaries | | | vowels | | | | updated V-UV-S | | label |
|----------|--------------------|-----|--------|--------|-----|-----|------|----------------|-----|-------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 1 | 19 | 19 | 0 | - | - | - | 4 | 6 | UP |
| 2 | 20 | 33 | 14 | 1 | 27 | 10 | /a/ | 1 | 1 | a |
| 3 | 34 | 62 | 29 | 0 | - | - | - | 1 | 1 | VC |
| 4 | 63 | 85 | 23 | 0 | - | - | - | 2 | 2 | UF |
| 5 | 86 | 101 | 16 | 1 | 93 | 10 | /a/ | 1 | 1 | a |
| 6 | 102 | 115 | 14 | 0 | - | - | - | 2 | 2 | UF |
| 7 | 116 | 127 | 12 | 0 | - | - | - | 1 | 1 | VC |
| 8 | 128 | 162 | 35 | 1 | 138 | 45 | /ii/ | 1 | 1 | ii |
| 9 | 163 | 170 | 8 | 0 | - | - | - | 1 | 1 | VC |
| 10 | 171 | 195 | 25 | 1 | 180 | 15 | /a/ | 1 | 1 | a |

Table 7.5 Segmentation results for the word 'tamsa^hiina'

According to a certain rule which is explained in the correction algorithms (Section 7.4.3), the voiced consonant preceding the vowel /ii/ has to be added to the vowel segment to yield the following string of labels:

UP-a-VC-UF-a-UF-ii-VC-a its syllabic pattern /CVC-CV-CVV-CV/

Now this word has only 9 segments, where segments 7 and 8 are combined together under the label 'ii', with a new length equal to 47 frames (12+35). The difference (VED-length) for this new vowel segment is now equal to -2 frames.

Another case of error is shown in the following example. Table 7.6 illustrates the segmentation results for the word 'ḍaraba'. These results are given graphically in Figure 7.9. This tables indicates that this word has four segments, and its string of labels is as follows:

VP-a-VP-a its syllabic pattern /CV-CV/

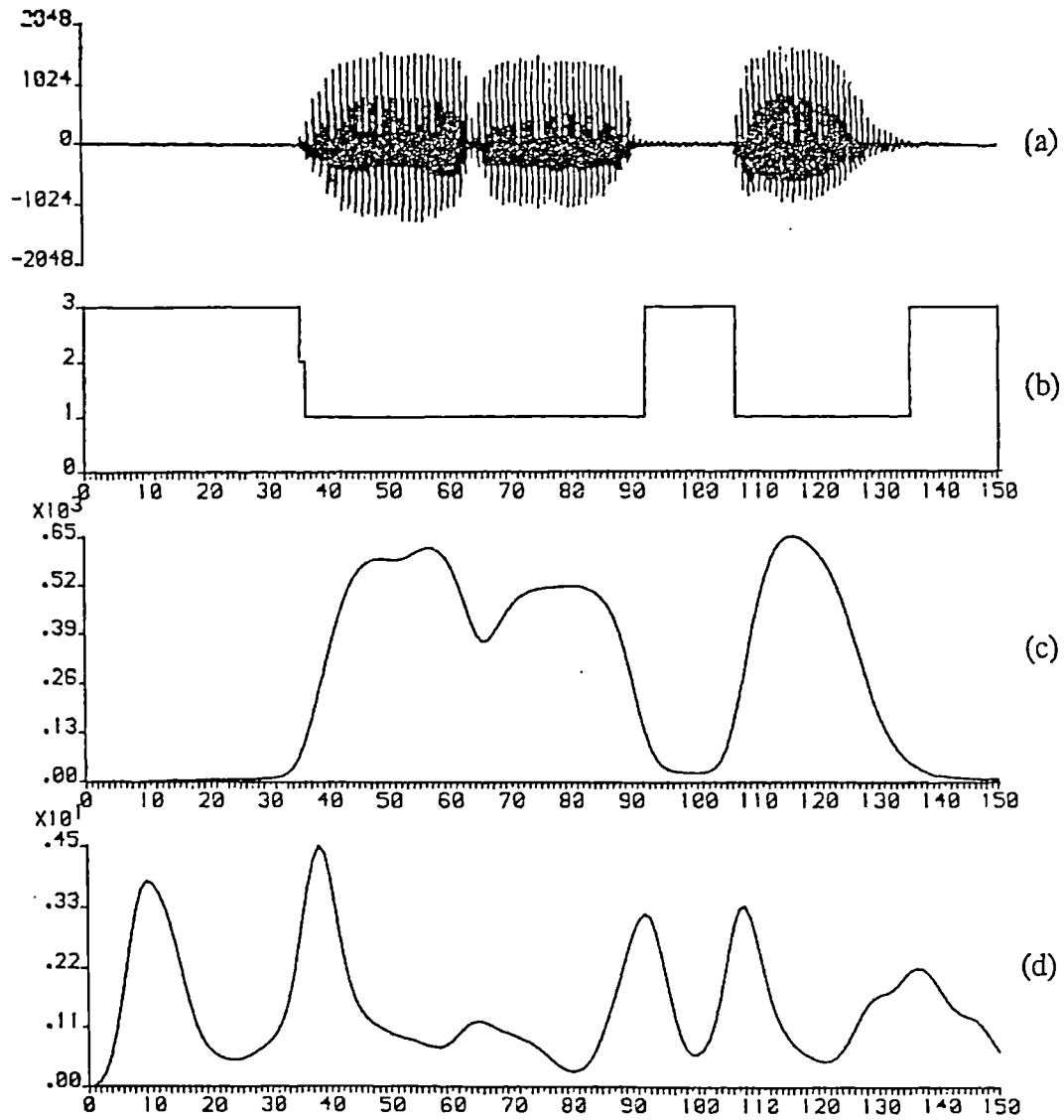


Fig. 7.9 Graphs for the word 'daraba'
 a) the speech signal, b) the V-UV-S contour
 c) the ES contour, d) the SV contour

| seg. no. | segment boundaries | | | vowels | | | | updated V-UV-S | | label |
|----------|--------------------|-----|--------|--------|--------|--------|----------|----------------|-----|-------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 11 | 38 | 28 | 0 | - | - | - | 5 | 6 | VP |
| 2 | 39 | 92 | 54 | 2 | 59, 81 | 21, 22 | /a/, /a/ | 1 | 1 | a |
| 3 | 93 | 109 | 17 | 0 | - | - | - | 5 | 5 | VP |
| 4 | 110 | 137 | 28 | 1 | 120 | 16 | /a/ | 1 | 1 | a |

Table 7.6 Segmentation results for the word 'daraba'

This pattern shows a legitimate syllabic structure, but this does not mean that the segmentation results are correct, (where the actual syllabic pattern is /CV-CV-CV/). It can be seen that the SV contour of this word (Figure 7.9d) displays a weak peak at frame 65, but this peak was discarded because its value is below the SVth (threshold of the SV contour) of this word. The ES contour of Figure 7.9c displays three peaks, and the vowel recognition procedure has given three vowels. According to the segmentation result table, both VRFs of the first and the second vowel (frames 59 and 81) lie within the boundaries of the second segment of this word. Therefore, Cd1 of the second segment in the result table (which refers to the presence of vowels) has been given the code '2' (during the labelling process), which refers to the presence of two vowels within this segment. The vowel representative frames and their estimated duration are also given in the neighbouring columns (this has been implemented by creating a second dimension in the result table or array).

However, the above mentioned case is dealt with in the vowel correction algorithm, where the second segment in Table 7.6 which contains two vowels is replaced by two vowel segments. Thus, the resultant string of labels for this word becomes:

VP-a-a-VP-a it syllabic pattern /CV-V-CV/

This pattern is illegitimate, and can be corrected by adding a voiced consonant between the first two vowels to yield the correct pattern:

VP-a-VC-a-VP-a it syllabic pattern /CV-CV-CV/

7.4 Error Correction Procedure

In the previous section, the types of errors which may occur during the segmentation and labelling process have been demonstrated. The aim of the error correction procedure is to overcome such errors which are caused by the limitation of the implemented algorithms (i.e., their impact on the initial segmentation results). The final segmentation results are expected to be an accurate estimation of the phonetic description of the speech signal at the input of the recognition system.

Figure 7.10 illustrates a block diagram of the implemented error correction procedure. The initial segmentation result table is provided at the input of the error correction procedure.

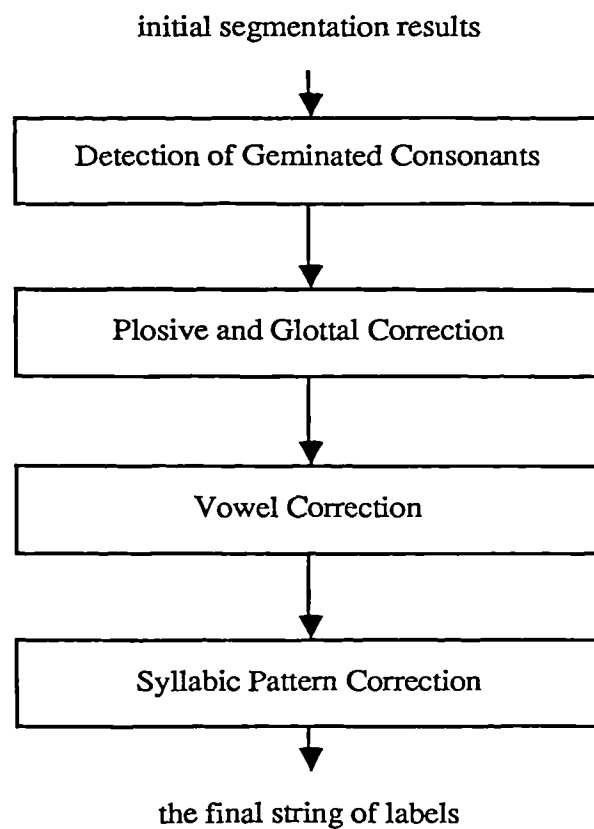


Fig. 7.10 A block diagram of the correction procedure

The first step in this procedure is to check the duration or length of each consonantal segment in the result table and to decide whether it belongs to a single or a geminated consonant. The second step, plosive correction, is carried out to overcome the case of having a plosive phoneme represented by separate silence and unvoiced (or burst) segments. Also, the presence of a glottal phoneme /ʔ/ at word-initial position is checked. The third step is to check the length of the vowel segments to decide whether each such segment comprises a vowel, a vowel and a voiced consonant, or part of a vowel, and to perform the necessary correction. The last step in the correction procedure is to check the syllabic pattern of a word and to adjust it to become a legitimate structure according to the phonological rules given in Section 3.7.2. The final string of labels is used for lexical access to locate the word (or the set of words) which shares the same labelling with the input word. The correction algorithms of Figure 7.10 are described in the following sections.

7.4.1 Detection of Geminated Consonants

In the implemented labelling scheme, consonants are given four labels, i.e., unvoiced plosive, voiced plosive, unvoiced fricative and voiced consonant. As explained in Section 3.7.3, geminated consonants occur only in word-medial positions, which means that the first and last consonant of any word must be a single consonant. Thus, a geminated consonant represents the leading consonant of the current syllable and the trailing consonant of the previous syllable. For example, the word 'yassaan' has two syllables /CVC-CVVC/, where the bold consonants refer to the geminated consonant /s/.

Figure 7.11 shows a flow chart of the geminated consonant detection and correction algorithm. This algorithm uses an array which contains the segmentation result table of the word under test, where the table contains N lines related to N segments. The algorithm checks the length and the label of each segment, (the chart of Figure 7.11 checks the geminated unvoiced fricative). If the segment length of certain consonants (i.e., 'UF', 'VC', 'UP', or 'VP') is above a certain threshold (LENth) the consonant is considered as a geminated consonant. In this case, a new segment is created following the geminated consonant segment, and this new segment has the same label as the previous segment. The length of the geminated segment is divided between the original and the new segment.

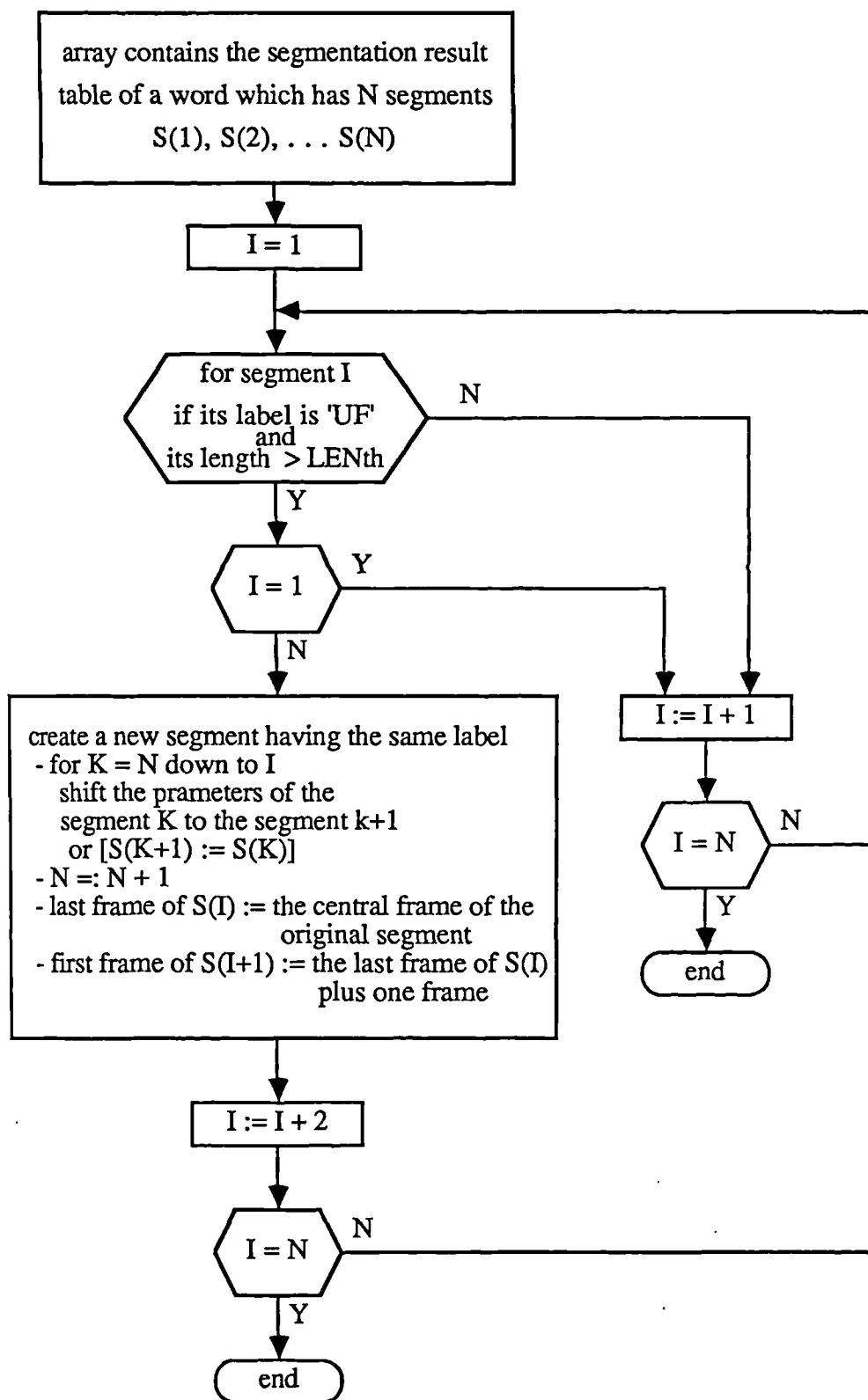


Fig. 7.11 Flow chart of geminated consonant detection and correction

For example, the segmentation results of the word 'yassaan' are given in Table 7.7, and these results are given graphically in Figure 7.12.

| seg. no. | segment boundaries | | | vowels | | | | updated V-UV-S | | label |
|----------|--------------------|-----|--------|--------|-----|-----|------|----------------|-----|-------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 14 | 23 | 10 | 0 | - | - | - | 1 | 1 | VC |
| 2 | 24 | 34 | 11 | 1 | 32 | 8 | /a/ | 1 | 1 | a |
| 3 | 35 | 78 | 44 | 0 | - | - | - | 2 | 2 | UF |
| 4 | 79 | 128 | 50 | 1 | 98 | 42 | /aa/ | 1 | 1 | aa |
| 5 | 129 | 143 | 15 | 0 | - | - | - | 1 | 1 | VC |

Table 7.7 The initial segmentation results for the word 'yassaan'

| seg. no. | segment boundaries | | | vowels | | | | updated V-UV-S | | label |
|----------|--------------------|-----|--------|--------|-----|-----|------|----------------|-----|-------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 14 | 23 | 10 | 0 | - | - | - | 1 | 1 | VC |
| 2 | 24 | 34 | 11 | 1 | 32 | 9 | /a/ | 1 | 1 | a |
| 3 | 35 | 56 | 22 | 0 | - | - | - | 2 | 2 | UF |
| 4 | 57 | 78 | 22 | 0 | - | - | - | 2 | 2 | UF |
| 5 | 79 | 128 | 50 | 1 | 98 | 44 | /aa/ | 1 | 1 | aa |
| 6 | 129 | 143 | 15 | 0 | - | - | - | 1 | 1 | VC |

Table 7.8 The modified segmentation results for the word 'yassaan'

The above algorithm is repeated three times to detect the presence of geminated voiced consonants, geminated unvoiced fricative consonants, and geminated plosive consonants in the word. In checking the geminated unvoiced fricative in the word 'yassaan' the third segment passes this test where its length is 44 frames, which is longer than the threshold LEN_{th} for unvoiced fricative consonants. Then the correction is made, and the modified results are given in Table 7.8.

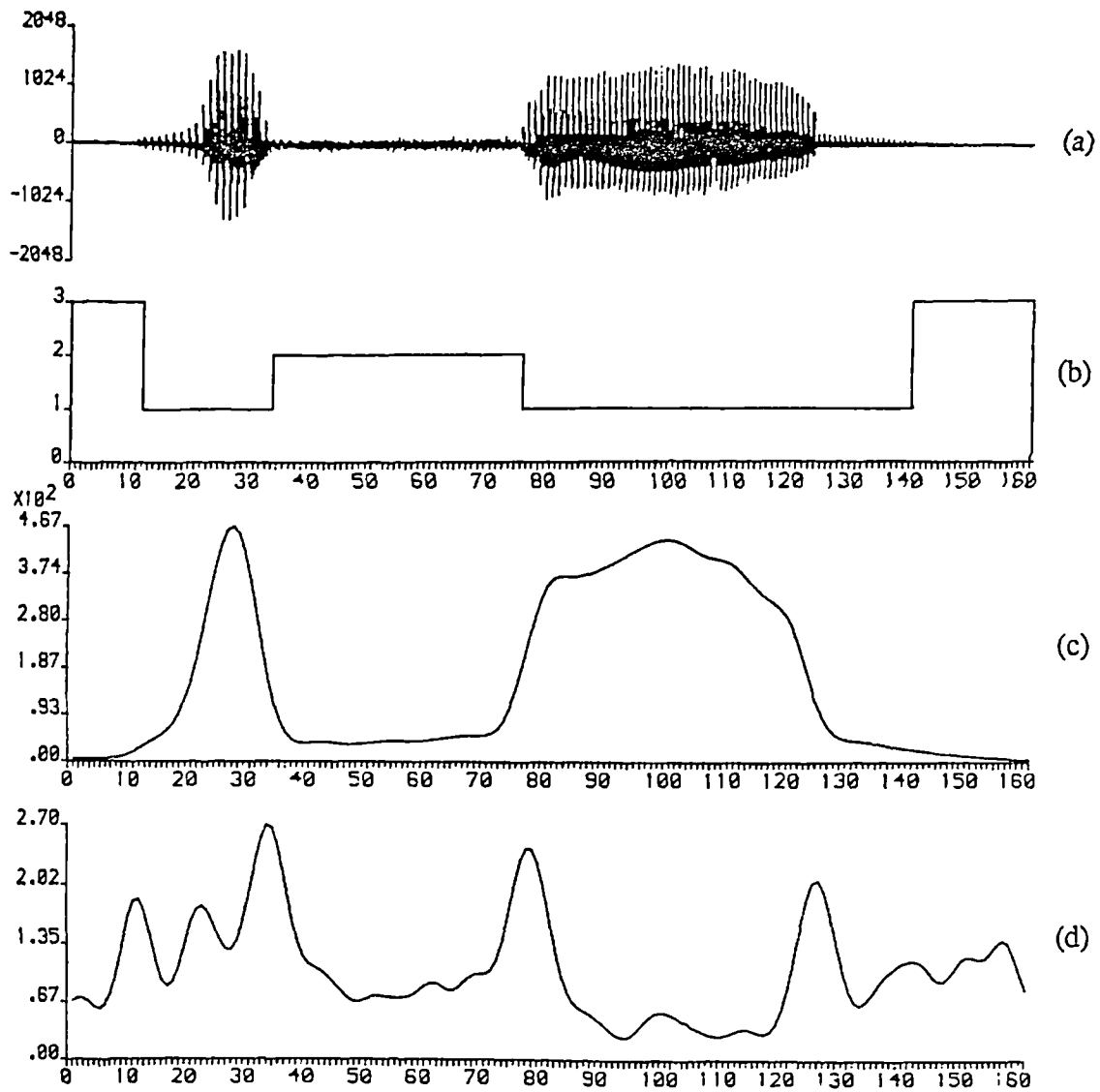


Fig. 7.12 Graphs for the word 'yassaan'
 a) the speech signal, b) the V-UV-S contour
 c) the ES contour, d) the SV contour

This correction algorithm is run three times, and each time is supplied with a different consonant label and the length threshold LEN_{th} for such label. LEN_{th} is taken equal to 30 frames (192 msec) for both 'VC' and 'UF', while it is taken equal to 20 frames (128 msec) for the plosive consonants (both 'VP' and 'UP'). Examples of other geminated consonants are given in Figures 6.8, 6.12, and 7.6. Figure 6.8 (in Chapter 6) shows graphs for the word 'jatasaffāhu' which contains the geminated unvoiced fricative /f/. Figure 6.12 shows graphs for the word 'laakinna' which contains the geminated consonant /n/. Figure 7.6 shows graphs for the word 'ʔaddiraasaati' which contains the geminated voiced plosive consonant /d/. The initial segmentation results of this word are given in Table 7.3.

7.4.2 Plosive and Glottal Correction

In general, the SV contour does not display a peak at the boundaries between the silence and unvoiced (burst) segments associated with the plosive phonemes. However, when such a peak occurs, both segments are combined together in one segment which maintains the label of the silence segment (i.e., either 'VP' or 'Up'). An example of such a case is given by the SV contour of the word 'ʔaddiraasaati' in Figure 7.4d, where it shows a peak between the silence and burst segments associated with the phoneme /t/. Table 7.3 shows that this peak is located at frame 192.

The glottal phoneme is categorised under the unvoiced plosive class (see Section 3.5.1). Many Arabic words start with the glottal phoneme. Normally, it is very difficult to detect the presence of this phoneme using a simple method. This phoneme is treated as an unvoiced plosive phoneme, where the V-UV-S contour of a certain word which starts with the glottal phoneme, may show a burst segment at the beginning, as in the word 'ʔaḵaama' given in Figure 7.13. Sometimes, the V-UV-S contour of such word shows no burst segment as in the word 'ʔafḍal' given in Figure 7.14. In the latter case, the presence of the glottal phoneme has to be estimated from the formant structures at the beginning of such a word.

However, a simple rule is applied in our system, that whenever a word starts with a vowel segment, a glottal or unvoiced plosive phoneme should be added before that vowel. This rule (or the whole plosive correction process) is applied after vowel correction, because the first vowel segment in a certain word might comprise a voiced consonant followed by a vowel.

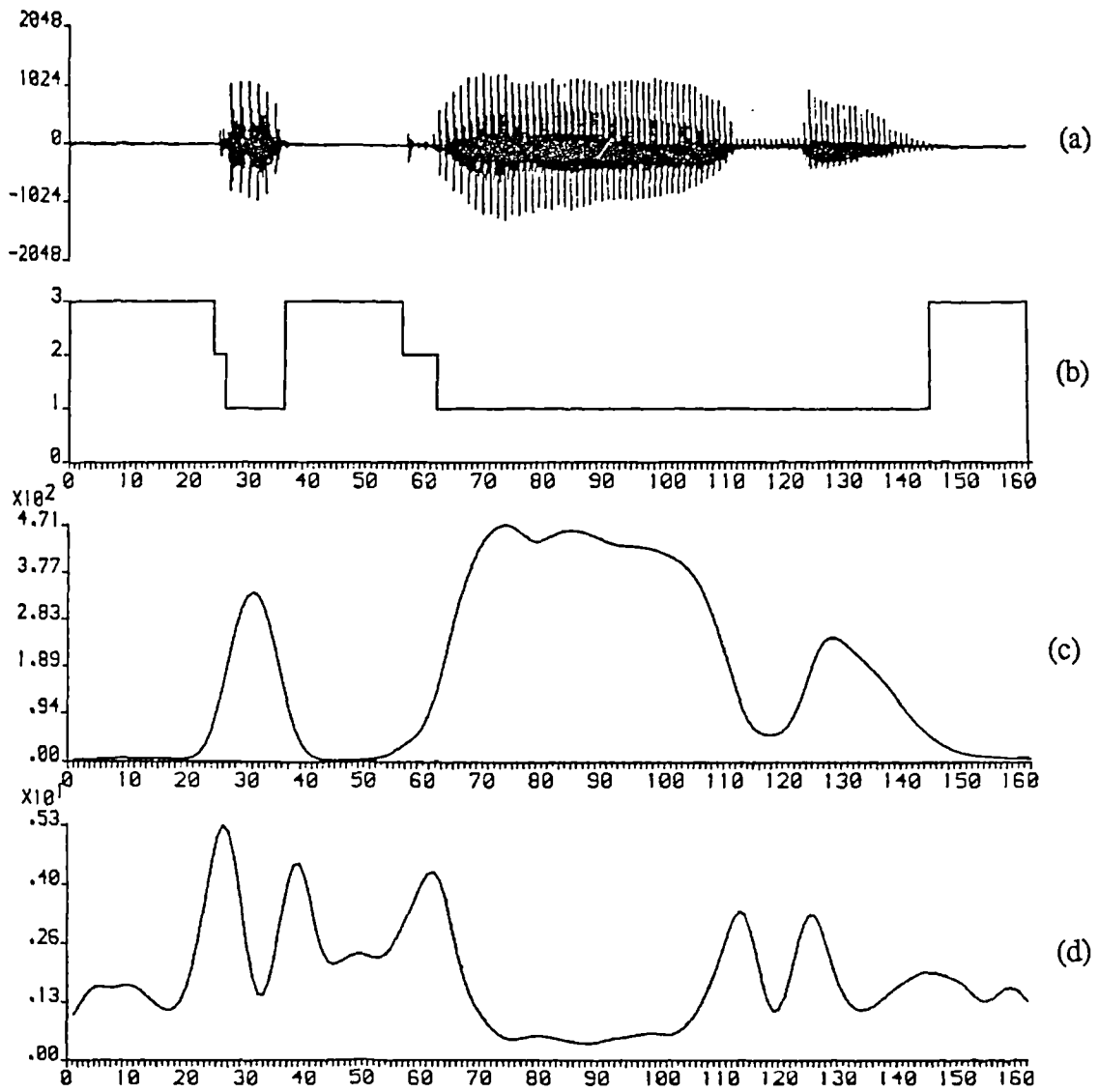


Fig. 7.13 Graphs for the word 'ʔakaama'
 a) the speech signal, b) the V-UV-S contour
 c) the ES contour, d) the SV contour

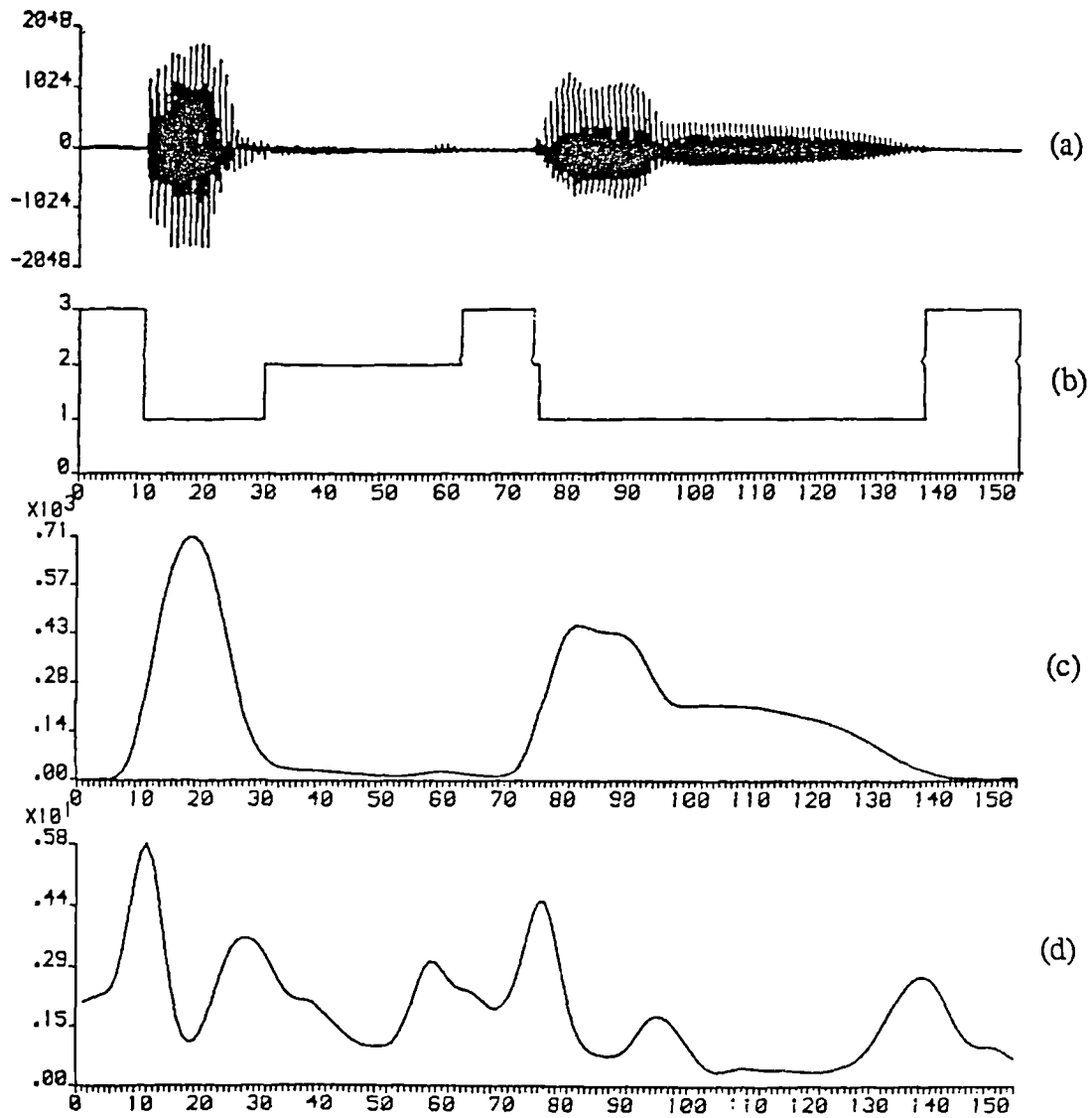


Fig. 7.14 Graphs for the word 'ʔafdal'
 a) the speech signal, b) the V-UV-S contour
 c) the ES contour, d) the SV contour

7.4.3 Vowel Correction

In the vowel correction procedure a comparison is made between the vowel estimated durations (VEDs) which are estimated from the ES contour and the length of the related vowel segment measured on the SV contour. Despite the title of this section, this algorithm handles only the following cases:

- VED of a certain vowel is longer than the length of its related segment on the SV contour (or $VED > \text{vowel segment length}$).
- VED of a certain vowel is shorter than the length of its related segment on the SV contour ($VED < \text{vowel segment length}$).
- Two vowel representative frames (VRFs) lie within the boundaries of one segment on the SV contour.

Other cases such as missing vowels are tackled in the syllabic pattern correction algorithm which is explained in the following sections.

Figure 7.15 illustrates the cases of errors which is tackled in this section. Figures 7.15a and 7.15b show the the case of having a VED longer than the vowel segment length. Figures 7.15c and 7.15d show the case of having a VED shorter than the vowel segment. Figure 7.15e shows the case where two vowel representative frames lie within the boundaries of the same vowel segment.

The vowel correction algorithm which deals with the first two cases is given in Figure 7.16. In this algorithm, the difference Δ between the vowel estimated duration (VED) of a certain vowel and the related vowel segment length (VSL) on the SV contour (which includes that vowel), is tested. When Δ is greater than a certain threshold (U_{th}) or it is less than another threshold (L_{th}), the vowel segment length is modified as explained in the following paragraphs.

When Δ is greater than U_{th} , we distinguish two cases. In the first case, the VED of the vowel under test extends over the vowel segment (which contains the VRF) and the following segment. This is shown in Figure 7.15a, where the vowel segment is bounded by P1 and P2 on the SV contour (the VRF lies between P1 and P2), and the VED [X(E)-X(B)] is longer than the duration [X(P2)-X(P1)] and expands over the next segment which is determined between P2 and P3. In the second case, the VED of the

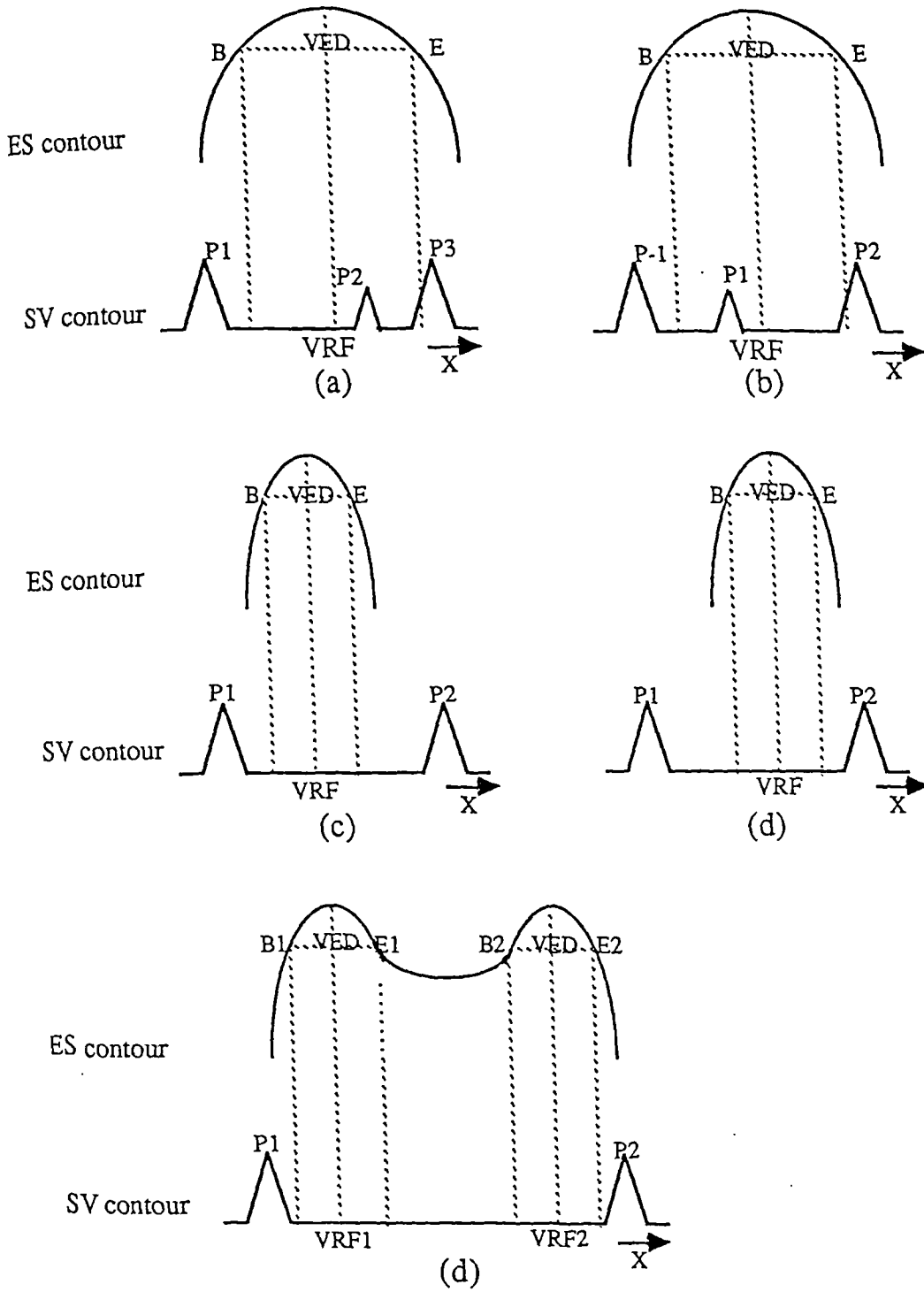


Fig. 7.15 Vowel errors

vowel under test extends over the vowel segment (which contains the VRF) and the preceding segment. This is shown in Figure 7.15b, where the vowel segment is bounded by P1 and P2 (the VRF lies between P1 and P2), and VED $[X(E) - X(B)]$ is longer than the duration $[X(P2) - X(P1)]$ and extends over the preceding segment between P-1 and P1. Thus, the boundaries of the vowel segment are modified by appending the following voiced segment in the first case, or the preceding voiced segment in the second case (see the algorithm in Figure 7.16).

When Δ is less than Lth, we also distinguish two cases. In the first case, the vowel segment contains the vowel and the following voiced segment. This is shown in Figure 7.15c, where the VRF lies in the left half of the vowel segment, and VED $[X(E) - X(B)]$ is shorter than the duration $[X(P2) - X(P1)]$. In the second case, the vowel segment contains the vowel and the preceding voiced segment. This is shown in Figure 7.15d, where the VRF lies in the right half of the vowel segment, and VED $[X(E) - X(B)]$ is shorter than the duration $[X(P2) - X(P1)]$. Thus, the boundaries of the vowel segment are modified by creating a new voiced segment which follows the vowel in the first case, or precedes the vowel in the second case (see the algorithm in Figure 7.16).

The case of having two vowels within one vowel segment on the SV contour (see Figure 7.15e), is tackled either by replacing this segment by two vowels, where the missing consonant between vowels is added during the syllabic pattern correction, or by replacing this segment by three segments, i.e., vowel, voiced consonant, and vowel. In the latter solution, the boundaries between the three segments are given as follows:

- the first vowel segment is taken between frames $X(P1)$ and $X(E1)$.
- the second vowel segment is taken between frames $X(B2)$ and $X(P2)$.
- the voiced consonant segment is taken between frames $X(E1)+1$ and $X(B2)-1$.

The values of the thresholds Lth and Uth are empirically chosen as follows:

- Uth is taken equal to 2 for short vowels, and equal to 5 for long vowels.
- Lth is taken equal -14 for vowels at word-final position, and equal to -7 elsewhere.

Finally, the segmentation result table is updated after each modification.

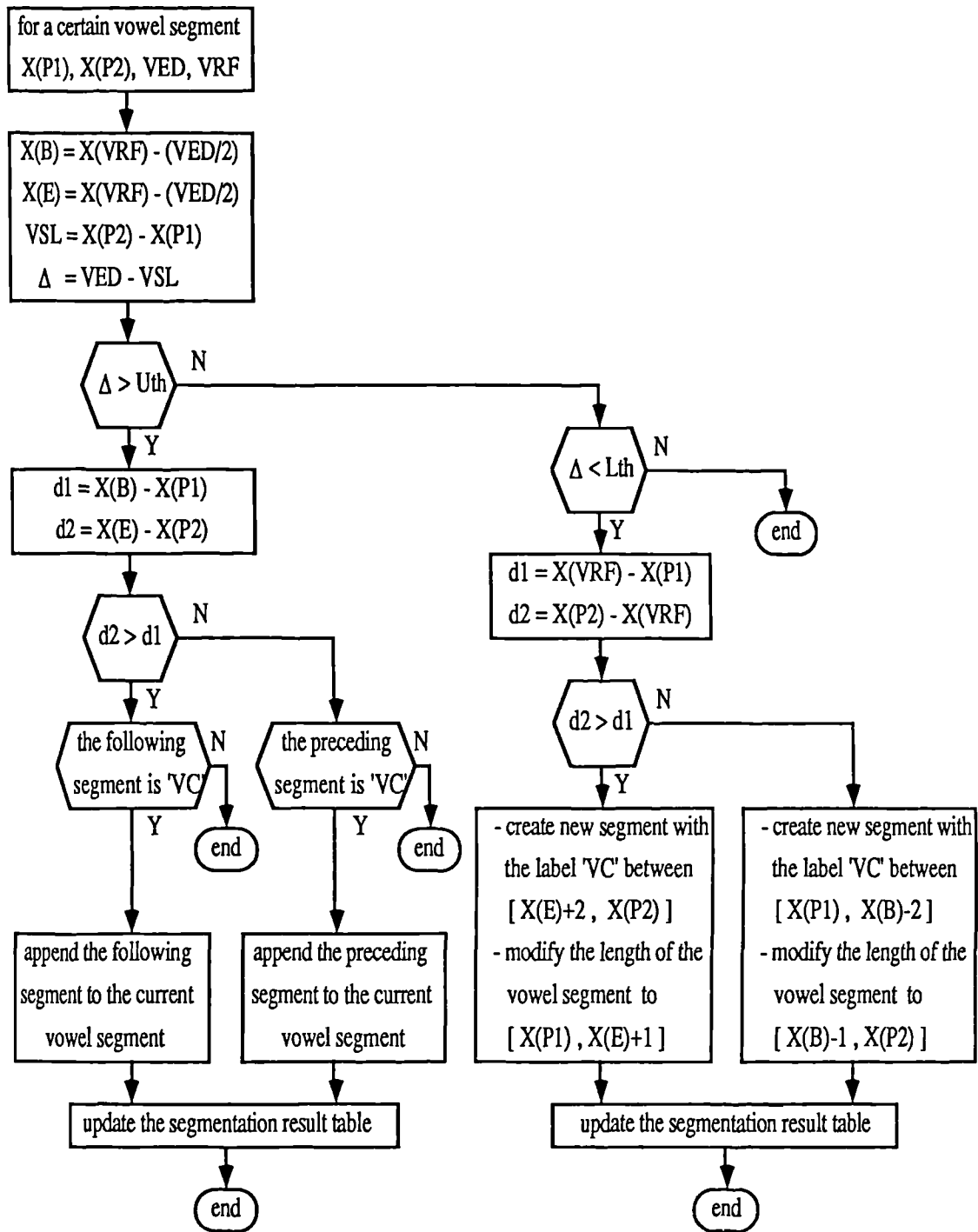


Fig. 7.16 Vowel correction algorithm

Figure 7.7d (the SV contour of the word 'jafealuuna') shows an example of VED shorter than the vowel segment length. The SV contour does not display a peak between the consonant /ε/ and the following vowel /a/. Thus, a voiced consonant is to be added before the vowel within the boundaries of the vowel segment (P4 , P5). Table 7.4 illustrates the initial segmentation results for this word. The length of the fourth segment (vowel segment 'a') in this table is 29 frames, while the VED of the vowel which lies within this segment is 13 frames. Therefore, a voiced consonant is to be created before the vowel, because the VRF (frame 93) lies in the left half of the vowel segment. The original vowel segment is replaced by the following two segments:

- 'VC' between frames 72-82
- 'a' between frames 83-100

Figure 7.8d (the SV contour of the word 'tamsaḥiina') shows an example of a VED longer than the vowel segment length. The SV contour has an extra peak along the vowel /ii/ and this has led to dividing the vowel into a voiced segment and a vowel segment. This voiced segment is then appended to the vowel segment. Table 7.5 illustrates the initial segmentation results for this word. The difference between the length of the eighth segment (vowel segment) in this table and the related vowel is 10 frames which is greater than Uth. Thus, the vowel segment should extend over the range 116-161 according to the VRF and the VED. The preceding voiced segment (the seventh segment in the table) is simply added to the vowel segment.

Table 7.9 shows the segmentation results for the word 'tamsaḥaani', and these results are given graphically in Figure 7.17. The seventh segment in this table is a vowel segment. The difference between the VED of the vowel which lies within this segment and its length is -14 frames, which is less than Lth. Thus, the vowel segment is split into two segments, a vowel segment followed by a voiced consonant segment. The vowel segment extends over the range 117-163 according to its VRF and VED, and the voiced consonant segment extends over the range 164-176 as illustrated in the modified result table (Table 7.10). Actually, the SV contour of this word (Figure 7.17d) shows a smoothed peak at frame 166 which is very near to the estimated boundaries of the new segment.

| seg. no. | segment boundaries | | | vowels | | | | updated V-UV-S | | label |
|----------|--------------------|-----|--------|--------|-----|-----|------|----------------|-----|-------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 1 | 20 | 20 | 0 | - | - | - | 4 | 6 | UP |
| 2 | 21 | 34 | 14 | 1 | 29 | 11 | /a/ | 1 | 1 | a |
| 3 | 35 | 62 | 28 | 0 | - | - | - | 1 | 1 | VC |
| 4 | 63 | 84 | 22 | 0 | - | - | - | 2 | 2 | UF |
| 5 | 85 | 99 | 15 | 1 | 90 | 10 | /a/ | 1 | 1 | a |
| 6 | 100 | 116 | 17 | 0 | - | - | - | 2 | 2 | UF |
| 7 | 117 | 175 | 59 | 1 | 139 | 45 | /aa/ | 1 | 1 | aa |
| 8 | 176 | 194 | 19 | 1 | 180 | 11 | /i/ | 1 | 1 | i |

Table 7.9 The initial segmentation results for the word 'tamsaḥaani'

| seg. no. | segment boundaries | | | vowels | | | | updated V-UV-S | | label |
|----------|--------------------|-----|--------|--------|-----|-----|------|----------------|-----|-------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 1 | 20 | 20 | 0 | - | - | - | 4 | 6 | UP |
| 2 | 21 | 34 | 14 | 1 | 29 | 11 | /a/ | 1 | 1 | a |
| 3 | 35 | 62 | 28 | 0 | - | - | - | 1 | 1 | VC |
| 4 | 63 | 84 | 22 | 0 | - | - | - | 2 | 2 | UF |
| 5 | 85 | 99 | 15 | 1 | 90 | 10 | /a/ | 1 | 1 | a |
| 6 | 100 | 116 | 17 | 0 | - | - | - | 2 | 2 | UF |
| 7 | 117 | 163 | 47 | 1 | 139 | 45 | /aa/ | 1 | 1 | aa |
| 8 | 164 | 176 | 13 | 0 | - | - | - | 1 | 1 | VC |
| 9 | 176 | 194 | 19 | 1 | 180 | 11 | /i/ | 1 | 1 | i |

Table 7.10 The modified segmentation results for the word 'tamsaḥaani'

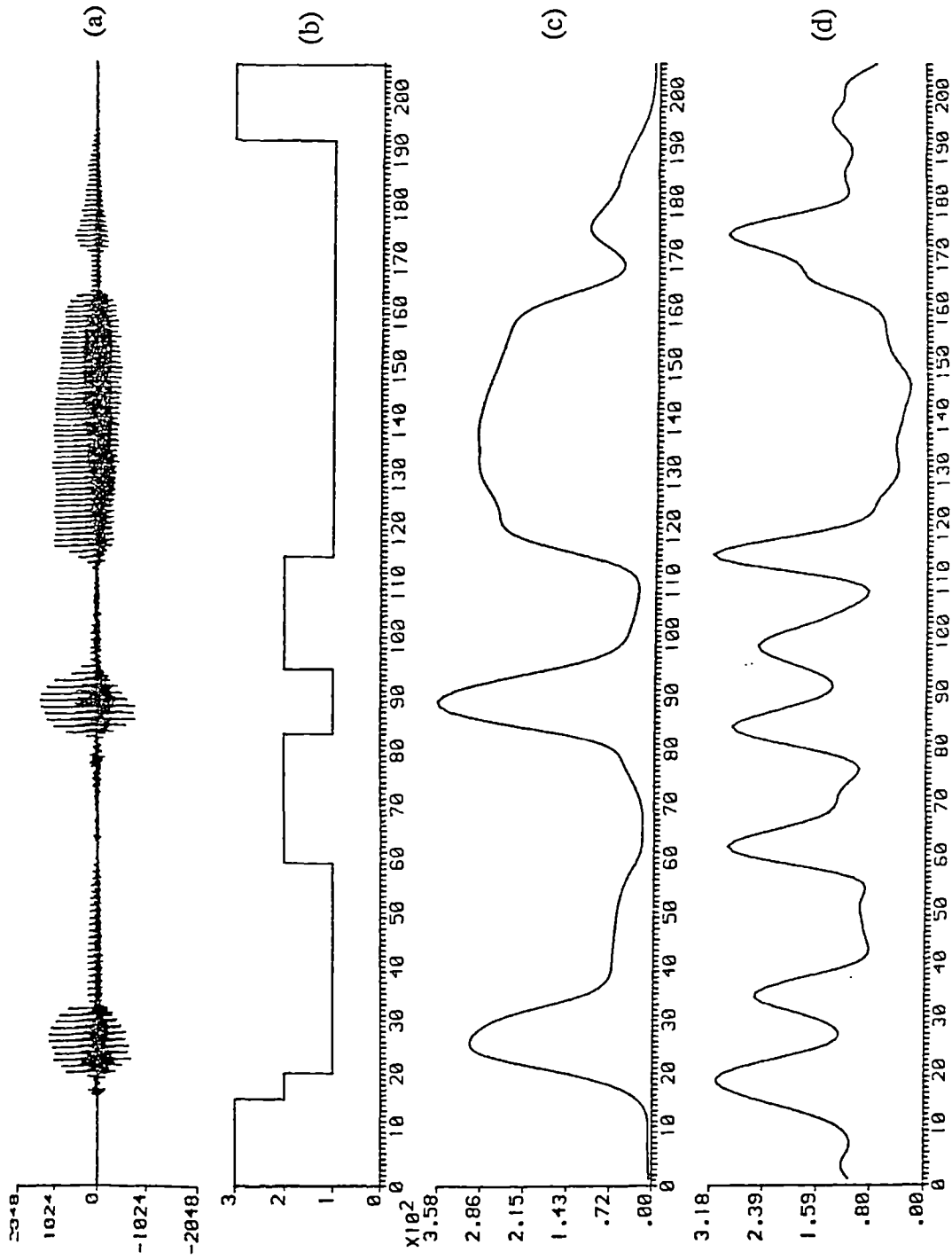


Fig. 7.17 Graphs for the word 'tamsahaani'
 a) the speech signal, b) the V-UV-S contour, c) the ES contour, d) the SV contour

7.4.4 Syllabic Pattern Correction

Up to this stage, the following corrections have been performed on the initial segmentation results:

- geminated consonant correction
- plosive correction
- vowel correction

The segmentation result table is updated after each correction process. We still have to tackle two more problems, which are the case of missing consonants and the case of having extra consonants even after vowel correction. The question which arises here, is how to decide that one of these problems is present. This is done by employing some phonological constraints (see Section 3.7.2).

The Arabic language uses five syllabic types (i.e., CV, CVV, CVC, CVVC, and CVCC) as explained in Section 3.7.1. Some of the phonological constraints (related to syllabic structures) are summarised as follows, (see Section 3.7.2):

- any syllable must have only one leading consonant
- the syllabic type /CVCC/ occurs only at word-final position
- the maximum allowable number of consonants in any consonant cluster in any word is two consonants
- the syllabic type /CVVC/, which has a long vowel, occurs mainly at word-final position, and it may occur at word-initial or word-medial position if its leading consonant is geminated

Thus, the syllabic structure or pattern of each word is checked to verify that it is not violating the above mentioned constraints or conditions. The following patterns show the likely error cases:

- a) a pattern starts with two consonants /CCV...../
- b) a pattern contains a cluster of three consonants /CVC-CCV...../
- c) a pattern of the form /...CVVC-CV...../
- d) a pattern ends with three consonants /...CVCCC/
- e) a pattern of the form /...CVVCC/
- f) a pattern of the form /...CV-VV...../

If any of the above errors occur in a certain syllabic pattern, it is considered as illegitimate pattern according to the above mentioned constraints.

a) The First Case

In this case, the syllabic pattern has two consonants in the leading consonant cluster of the first syllable. For example, assume that the segmentation and labelling process for a certain word has led to the following string of labels:

VC-VC-a-UF-aa-VC its syllabic pattern /CCV-CVVC/

This pattern is illegitimate and the first two voiced consonants must be combined together to form one segment. In this example these two consonants have the same labels (i.e., 'VC' and 'VC'). But if these consonants have different labels (e.g., 'UF' and 'VC'), the label belonging to the longer segment is maintained over the new segment. This latter case may occur when a word starts with a fricative consonant such as /ʒ/. This consonant is of a mixed excitation nature, therefore it could be voiced, unvoiced or mixed according to its position along a word. Table 7.11 illustrates the segmentation results for the word /ʒanuub/ (after applying plosive correction), and Figure 7.18 shows these results graphically. The V-UV-S contour (Figure 7.18b) shows two voiced and unvoiced segments related to the first consonant /ʒ/, and their lengths are given in the result table. Also the SV contour of this word (Figure 7.18d) shows two segments related to this consonant. The resultant syllabic pattern is /CCV-CVVC/ which is illegitimate. The result table shows that the unvoiced segment of this consonant is longer than the voiced segment. Thus, both the segments are combined together in one new segment, which is labelled as 'UF'. It is preferable to maintain an indication that such a case has occurred which may be useful in the verification stage of the speech recognition system (see Figure 4.4).

b) The Second Case

In this case, a cluster of three consonants occurs at word-medial position /...VCCCV.../. Then, the two segments which have the same label are combined together in one segment. For example, Table 7.12 illustrates the segmentation results for the word 'xamsa', where these results are given graphically in Figure 7.19.

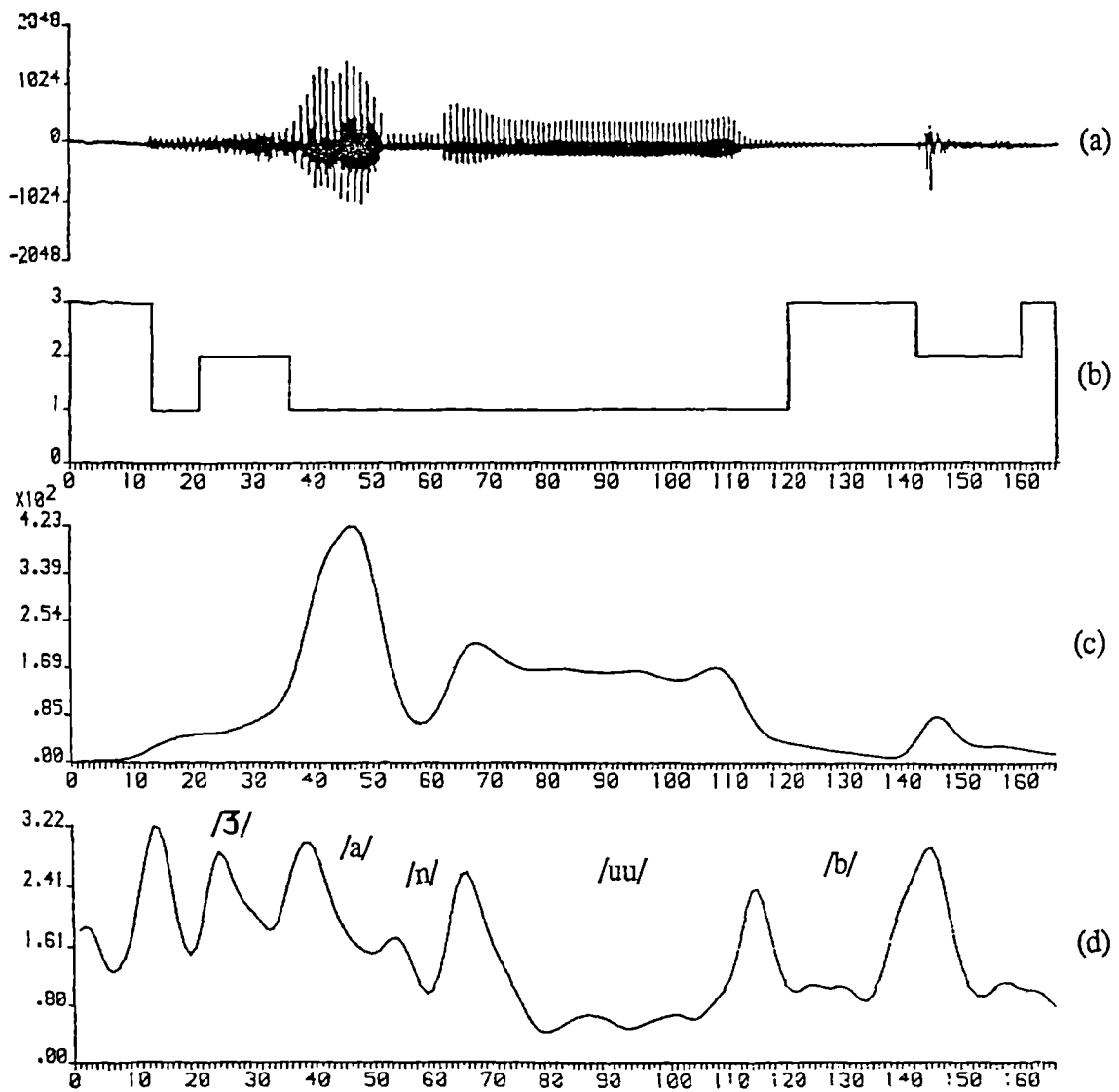


Fig. 7.18 Graphs for the word 'ʒanuub'
 a) the speech signal, b) the V-UV-S contour
 c) the ES contour, d) the SV contour

| seg. no. | segment boundaries | | | vowels | | | | updated V-UV-S | | label |
|----------|--------------------|-----|--------|--------|-----|-----|------|----------------|-----|-------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 15 | 24 | 10 | 0 | - | - | - | 1 | 1 | VC |
| 2 | 25 | 39 | 15 | 0 | - | - | - | 2 | 2 | UF |
| 3 | 40 | 53 | 14 | 1 | 47 | 11 | /a/ | 1 | 1 | a |
| 4 | 54 | 66 | 13 | 0 | - | - | - | 1 | 1 | VC |
| 5 | 67 | 116 | 50 | 1 | 89 | 49 | /uu/ | 1 | 1 | uu |
| 6 | 117 | 162 | 46 | 0 | - | - | - | 4 | 6 | Vp |

Table 7.11 The segmentation results for the word '3anuub'

| seg. no. | segment boundaries | | | vowels | | | | updated V-UV-S | | label |
|----------|--------------------|-----|--------|--------|-----|-----|-----|----------------|-----|-------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 16 | 27 | 12 | 0 | - | - | - | 2 | 2 | UF |
| 2 | 28 | 37 | 10 | 0 | - | - | - | 2 | 2 | UF |
| 3 | 38 | 52 | 15 | 1 | 45 | 12 | /a/ | 1 | 1 | a |
| 4 | 53 | 79 | 27 | 0 | - | - | - | 1 | 1 | VC |
| 5 | 80 | 88 | 9 | 0 | - | - | - | 2 | 2 | UF |
| 6 | 89 | 102 | 14 | 0 | - | - | - | 2 | 2 | UF |
| 7 | 103 | 127 | 25 | 1 | 113 | 15 | /a/ | 1 | 1 | a |

Table 7.12 The initial segmentation results for the word 'xamsa'

The SV contour of this word (Figure 7.19d) shows two segments for the first phoneme /x/ and also for the fourth phoneme /s/ in this word. This is shown in Table 7.12 as well, where the first and the second 'UF' segments belong to the first phoneme /x/, while the fifth and the sixth 'UF' segments belong to the phoneme /s/.

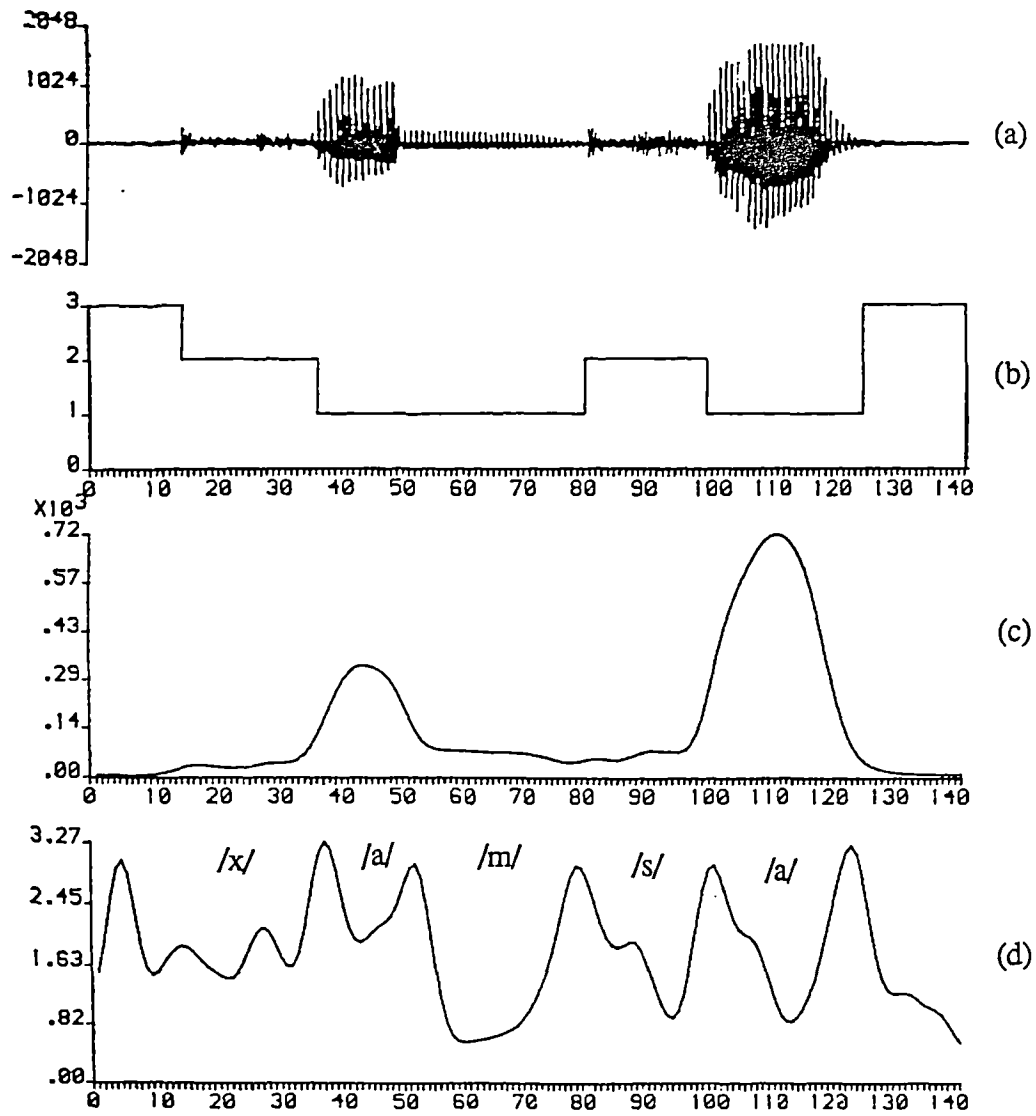


Fig. 7.19 Graphs for the word 'xamsa'
 a) the speech signal, b) the V-UV-S contour
 c) the ES contour, d) the SV contour

The resultant syllabic pattern is /CCVC-CCV/ which is an illegitimate pattern. This pattern can be corrected by combining the first and the second segment of Table 7.12 into one new segment which is labelled as 'UF', and combining the fifth and the sixth segments together into one segment which is labelled as 'UF'. The final string of labels is:

UF-a-VC-UF-a its syllabic pattern /CVC-CV/

c) The Third Case

In this case, the syllable /CVVC/ occurs at word-initial or word-medial positions and its trailing consonant is not geminated, i.e., the pattern /CVVC-CV...../ is not a legitimate pattern unless the two bold consonants are related to one geminated consonant. If they are not geminated, they are combined into one consonant which is the leading consonant of the second syllable. This case has not been encountered in our speech database.

d) The Fourth Case

In this case, the syllabic pattern of a certain word has three consonants in the trailing consonant cluster of the last syllable, such as /...CVCCC/. In this respect, the three consonants are combined together if they share the same label to yield the pattern /...CVC/. But if they are not the same, the two similar consonants are combined together and the other one is left as it is, therefore the resultant pattern is /...CVCC/. The latter case has not been encountered in our speech database.

e) The Fifth Case

In this case (as in the pattern /...CV-VV.../), the leading consonant of the second syllable is missing. This consonant may have not been recovered by the vowel correction algorithm due to its very short length. In this case, a voiced consonant should be created before the second vowel to yield the pattern /....CV-CVV.../.

f) The Sixth Case

In this case (as in the pattern /...CVVCC/), the last syllable has a long vowel, therefore the syllable is not allowed to have more than one trailing consonant. In this respect, three cases are distinguished:

- both consonants are unvoiced fricative, then they are grouped together under one 'VC' label
- both consonants are voiced consonants and their combined length is less than 30 frames, then they are grouped together under one 'VC' label
- both consonants are voiced consonants and their combined length is more than 30 frames. Then, the second consonant is relabelled as a vowel, and its identity can be found by calling the vowel identification procedure (the VRF is taken at the centre of that segment)

For example, the word 'jusaawii' has three vowels. The vowel detection procedure failed to detect the third vowel /ii/ (see Section 5.2.3), as illustrated in Figure 5.13 (in Chapter 5). This is demonstrated in the segmentation result table (Table 7.13) of this word. In this table, the last segment which is supposed to be the last vowel segment has been labelled as a voiced consonant segment 'VC', where the code Cd1 is '0' for this segment. The resultant syllabic pattern of this word is /CV-CVVCC/, which is an illegitimate pattern. Both consonants at the end of this word are voiced consonants, and their combined length is 69 frames. According to the above conditions, the second consonants is relabelled as a vowel, where its identity (the vowel /i/) is found by passing frame number 147 to the vowel identification procedure. The length of this vowel segment is 53 frames, therefore the vowel is labelled as the long vowel 'ii'. Thus, the resultant string of labels is

VC-u-UF-aa-VC-ii

its syllabic pattern

/CV-CVV-CVV/

| seg. no. | segment boundaries | | | vowels | | | | updated V-UV-S | | label |
|----------|--------------------|-----|--------|--------|-----|-----|------|----------------|-----|-------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 10 | 27 | 18 | 0 | - | - | - | 1 | 1 | VC |
| 2 | 28 | 41 | 14 | 1 | 33 | 14 | /a/ | 1 | 1 | u |
| 3 | 42 | 62 | 21 | 0 | - | - | - | 2 | 2 | UF |
| 4 | 63 | 104 | 42 | 1 | 85 | 39 | /aa/ | 1 | 1 | aa |
| 5 | 105 | 120 | 16 | 0 | - | - | - | 1 | 1 | VC |
| 6 | 121 | 173 | 53 | 0 | - | - | - | 1 | 1 | VC |

Table 7.13 The initial segmentation results for the word 'jusaawii'

Actually, the case of having missing vowels is very difficult to tackle. Sometimes the missing vowel can not be recovered, and the correction procedure may lead to modified phonetic descriptions of certain words. For example, the segmentation results of the word 'ʔarbaea' are given in Table 7.14, and these results are given graphically in Figure 7.20.

| seg. no. | segment boundaries | | | vowels | | | | updated V-UV-S | | label |
|----------|--------------------|-----|--------|--------|-----|-----|-----|----------------|-----|-------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 9 | 24 | 16 | 1 | 17 | 12 | /a/ | 1 | 1 | a |
| 2 | 25 | 36 | 12 | 0 | - | - | - | 1 | 1 | VC |
| 3 | 37 | 49 | 13 | 0 | - | - | - | 5 | 5 | VP |
| 4 | 50 | 99 | 50 | 1 | 61 | 18 | /a/ | 1 | 1 | a |

Table 7.14 The initial segmentation results for the word 'ʔarbaea'

| seg. no. | segment boundaries | | | vowels | | | | updated V-UV-S | | label |
|----------|--------------------|-----|--------|--------|-----|-----|-----|----------------|-----|-------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 1 | 8 | 7 | 0 | - | - | - | 4 | 4 | UP |
| 2 | 9 | 24 | 16 | 1 | 17 | 12 | /a/ | 1 | 1 | a |
| 3 | 25 | 36 | 12 | 0 | - | - | - | 1 | 1 | VC |
| 4 | 37 | 49 | 13 | 0 | - | - | - | 5 | 5 | VP |
| 5 | 50 | 70 | 21 | 1 | 61 | 18 | /a/ | 1 | 1 | a |
| 6 | 71 | 99 | 29 | 0 | - | - | - | 1 | 1 | VC |

Table 7.15 The modified segmentation results for the word 'ʔarbaea'

The initial syllabic pattern is /VC-CV/. Then during the correction procedure, an unvoiced plosive label is added before the first vowel by the plosive correction algorithm, and the vowel correction algorithm divides the last vowel into vowel and voiced consonant as

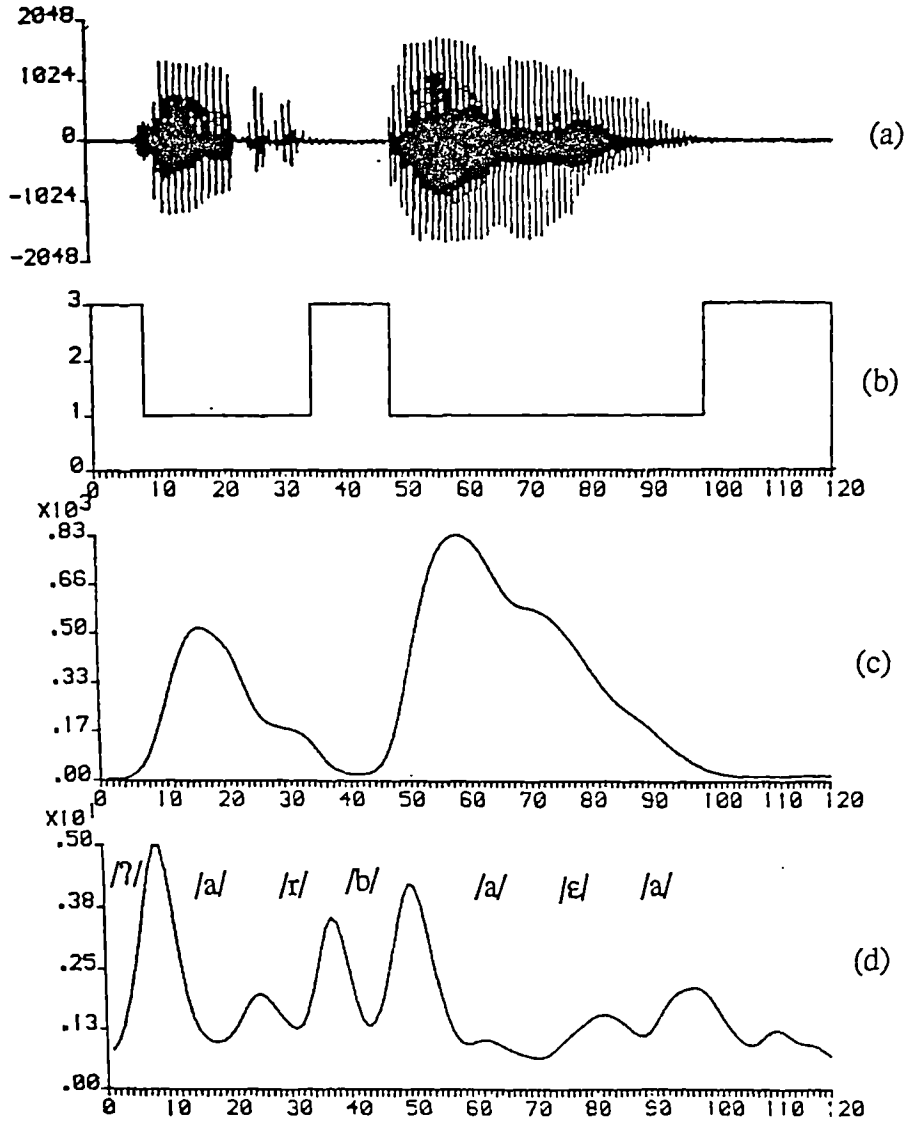


Fig. 7.20 Graphs for the word 'arbaea'
 a) the speech signal, b) the V-UV-S contour
 c) the ES contour, d) the SV contour

illustrated in the modified results given in Table 7.15. From this table, the resultant syllabic pattern is /CVC-CVC/. As we notice, the vowel and the consonant of the last syllable in the word 'ʔarbaea' (i.e., /εa/) have been combined into one voiced consonant segment, and the correction procedure has failed to recover the missing vowel. Note that the SV contour of this word (Figure 7.20d) displays a low peak at frame 84, which is neglected because its value falls below the SVth of this word. If this peak is detected as a valid peak, the resultant syllabic pattern would be /CVC-CVC/, and after vowel correction it would become /CVC-CVCC/. Nevertheless, the case of a missing vowel at word-final position still requires more study.

Another example is the word 'sabea'. The segmentation results of this word are shown in Table 7.16, and these results are shown graphically in Figure 7.21. Table 7.17 illustrates the modified results after vowel correction. In this case, the last vowel segment in Table 7.16 is divided into a vowel segment and a voiced consonant segment as shown in Table 7.17. Then, the resultant syllabic structure is /CV-CVC/, while the actual syllabic structure of this word is /CVC-CV/. This difference occurs because of the consonant /ε/ whose energy is relatively higher than the following vowel /a/. It can be noticed that both words 'ʔarbaea', and 'sabea' have the voiced consonant /ε/ as the leading consonant of the last syllable. The presence of this consonant leads to syllabic patterns which are different from the actual patterns of both words. Such cases (i.e., the presence of the consonant /ε/), and other similar cases, can be taken into consideration in the lexicon, where a different phonetic description of such words could be used.

| seg. no. | segment boundaries | | | vowels | | | | updated V-UV-S | | label |
|----------|--------------------|-----|--------|--------|-----|-----|-----|----------------|-----|-------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 19 | 44 | 26 | 0 | - | - | - | 2 | 2 | UF |
| 2 | 45 | 61 | 17 | 1 | 57 | 11 | /a/ | 1 | 1 | a |
| 3 | 62 | 76 | 15 | 0 | - | - | - | 5 | 6 | VP |
| 4 | 77 | 129 | 53 | 1 | 89 | 17 | /a/ | 1 | 1 | a |

Table 7.16 The initial segmentation results for the word 'sabea'

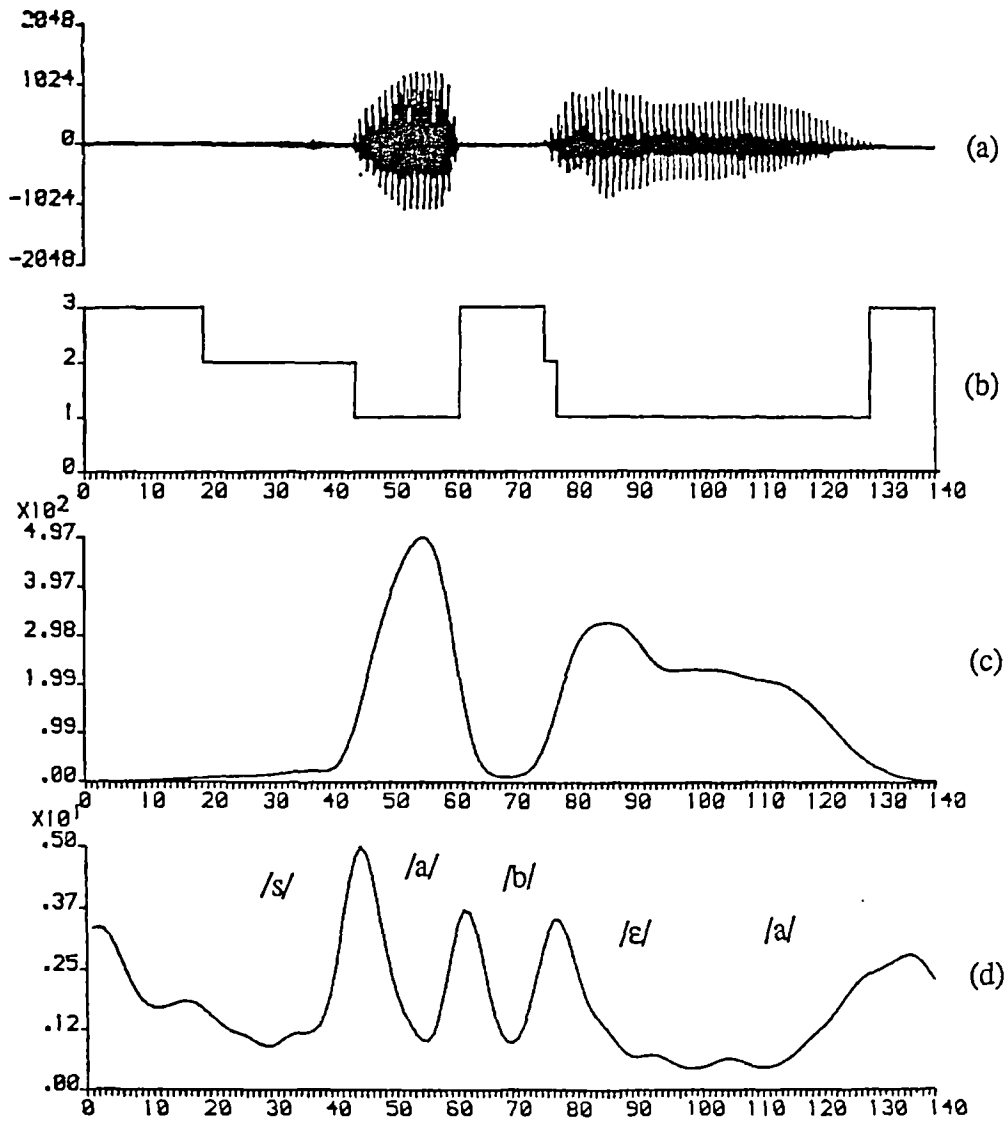


Fig. 7.21 Graphs for the word 'sabæa'
 a) the speech signal, b) the V-UV-S contour
 c) the ES contour, d) the SV contour

| seg. no. | segment boundaries | | | vowels | | | | updated V-UV-S | | label |
|----------|--------------------|-----|--------|--------|-----|-----|-----|----------------|-----|-------|
| | begin | end | length | Cd1 | VRF | VED | ID | Cd2 | Cd3 | |
| 1 | 19 | 44 | 25 | 0 | - | - | - | 2 | 2 | UF |
| 2 | 45 | 61 | 17 | 1 | 57 | 11 | /a/ | 1 | 1 | a |
| 3 | 62 | 76 | 15 | 0 | - | - | - | 5 | 6 | VP |
| 4 | 77 | 94 | 18 | 1 | 89 | 17 | /a/ | 1 | 1 | a |
| 5 | 95 | 129 | 35 | 0 | - | - | - | 1 | 1 | VC |

Table 7.17 The modified segmentation results for the word 'sabea'

7.5 Results and Discussion

The segmentation and error correction algorithms have been applied to a set of 570 words. These words comprise the repetition of the basic one hundred words uttered by 5 speakers (4 males, and 1 female) as described in Section 5.2.3. The basic set of 100 words contains monosyllabic and polysyllabic words (see Appendix A).

Table 7.18 illustrates the syllabic structures which are found in this set. The first column displays the number of syllables in each pattern. The second column shows the syllabic structures. The third column gives the number of words sharing the same syllabic structure. The fourth column gives the number of different morphological balances under each syllabic structure, while column five shows the number of different patterns or strings of labels according to the implemented classification scheme (scheme number 7, see Section 4.3.1). It can be seen from this table, that the basic set of 100 words contains 4 monosyllabic words, 39 disyllabic words, 41 trisyllabic words, 13 quadrisyllabic words, and three five-syllabic words. The table shows also that this set has 30 different syllabic structures, 50 different morphological balances, and 96 different strings of labels or patterns according to the used classification scheme. These 30 different syllabic structures are about 42% of those structures (72 structures) found in the 10,000 words lexical database (see Section 4.3.2).

The error rate made by each speaker is given in Table 7.19, where the average error rate for the 570 test words is about 5%. These errors are mainly originated by the vowel

| number of syllables | syllabic structure | number of words | number of balances | number of patterns |
|---------------------------|--------------------|-----------------------|--------------------------|--------------------------|
| 1 | CVV | 2 | 2 | 2 |
| 1 | CVC | 1 | 1 | 1 |
| 1 | CVCC | 1 | 1 | 1 |
| 2 | CV-CV | 1 | 1 | 1 |
| 2 | CV-CVV | 2 | 2 | 2 |
| 2 | CV-CVC | 1 | 1 | 1 |
| 2 | CV-CVVC | 8 | 4 | 7 |
| 2 | CVV-CV | 5 | 1 | 4 |
| 2 | CVV-CVC | 4 | 1 | 4 |
| 2 | CVC-CV | 4 | 2 | 4 |
| 2 | CVC-CVV | 2 | 1 | 2 |
| 2 | CVC-CVC | 8 | 2 | 8 |
| 2 | CVC-CVVC | 4 | 3 | 4 |
| 3 | CV-CV-CV | 18 | 1 | 16 |
| 3 | CV-CVV-CV | 6 | 3 | 6 |
| 3 | CV-CVV-CVV | 1 | 1 | 1 |
| 3 | CVV-CVV-CVV | 1 | 1 | 1 |
| 3 | CVV-CVC-CV | 1 | 1 | 1 |
| 3 | CVC-CV-CV | 10 | 5 | 10 |
| 3 | CVC-CV-CVV | 2 | 2 | 2 |
| 3 | CVC-CVV-CV | 2 | 1 | 2 |
| 4 | CV-CVV-CV-CV | 2 | 2 | 2 |
| 4 | CV-CVV-CV-CVV | 1 | 1 | 1 |
| 4 | CVV-CV-CV-CVV | 1 | 1 | 1 |
| 4 | CVC-CV-CVV-CV | 7 | 4 | 7 |
| 4 | CVC-CV-CV-CVC | 1 | 1 | 1 |
| 4 | CVC-CV-CV-CV | 1 | 1 | 1 |
| 5 | CV-CV-CVC-CV-CV | 1 | 1 | 1 |
| 5 | CVC-CVC-CV-CV-CV | 1 | 1 | 1 |
| 5 | CVC-CV-CVV-CVV-CV | 1 | 1 | 1 |

Table 7.18 syllabic patterns for the test set

detection procedure. As it has been shown in Section 5.3, the vowel detection procedure leads to two types of errors.

| speaker | MZ | YM | HT | MB | HK |
|--------------|-----|-----|-----|----|----|
| no. of words | 200 | 200 | 100 | 50 | 20 |
| error rate | 3% | 7% | 5% | 6% | 5% |

Table 7.19 Error rates

The first type is the presence of weak peaks along the ES contour related to some weak vowels. For example, for the words 'ʔaddiraasaati' and 'zaaʔid' uttered by speaker YM, the peaks related to the vowel /i/ in the syllables /di/ and /ʔid/ have been ignored because they were very weak and their values fell below the thresholds (ES_{th} for each word) of the vowel detection algorithm. Such cases have not been seen in the realisation of the same words by other speakers.

The second type of error is the case of having a missing vowel peak and/or having a vowel peak within the boundaries of the preceding voiced consonant segment. These two cases are clearly shown in the previous section for the word 'ʔarbaea' (Figure 7.18) and for the word 'sabea' (Figure 7.19). The ES contour of the first word (see Figure 7.18c) shows no peak related to the third vowel, because of the voiced consonant /ε/. So, the resultant string of labels for this word is /UP-a-VC-VP-a-VC/ where the actual string of labels is /UP-a-VC-VP-a-VC-a/. In the second word 'sabea', the ES contour (Figure 7.19c) shows a peak within the boundaries of the voiced consonant segment /ε/. Thus, the segmentation result is the following string of labels /UF-a-VP-a-VC/ where the actual string of labels is /UF-a-VP-VC-a/. These two cases, besides similar cases, must be taken into consideration by altering the phonetic description of these words in the lexicon.

Another source of errors is the different realisation of consonants, especially at word-final position. For example, the words 'waahid' and 'zaaʔid' are terminated by the voiced plosive phoneme /d/. Such phoneme at word-final position is in free variation, i.e., voiced or unvoiced, released or unreleased (pronounced or not pronounced). The phoneme /d/ at word-final position is mostly unreleased. Another example is the word

'sifr', where the phoneme /r/ is mostly unvoiced when it occurs in word-final position. Such cases can be dealt with when building the lexicon by having multiple entries in the lexicon for the same word.

As a result, the accuracy of the segmentation and labelling procedure depends heavily on the accuracy of the vowel detection procedure.

The result of the segmentation process for each word is a string of labels, which is used later on for lexical access to retrieve the word or the set of words sharing the same phonetic description. Also, the syllabic structure and the morphological balance of the test-word is determined at the end of the segmentation and labelling procedure. These results are very useful in speech understanding systems.

Finally, the set of 100 words used to test the system has 96 different patterns according to the implemented classification scheme (scheme number 7, see Section 4.3.1). The statistical results given in Section 4.3.2 have shown that 53.65% of the 10,000 words in the lexical database, are uniquely represented by this scheme, where the maximum number of words sharing the same string of labels is 17. Thus, these results show that even with this relatively simple classification scheme the number of words to be handled in the detailed verifier stage of the the proposed recognition system (see Section 4.5), is relatively small.

7.6 Summary

The segmentation and error correction procedures have been demonstrated in this chapter. The results of the voiced-unvoiced-silence segmentation procedure (given in Chapter 5), have been modified in this chapter to account for the presence of both voiced and unvoiced plosive phonemes.

The results of the modified voiced-unvoiced-silence segmentation procedure, the vowel recognition procedure, and the spectral variation function of each word (given in Chapter 6), have been employed in the labelling and error correction processes. The error correction procedure has employed durational information (prosodic feature) in addition to some phonological constraints of the Arabic language. The accuracy of segmentation results according to the relatively simple classification scheme is about 95%. This accuracy can be improved by improving the accuracy of the vowel detection procedure.

Chapter 8

Conclusions and Suggestions for Further Work

8.1 Introduction

This chapter provides an overview of the work as a whole and refers to the most important results. Also, it highlights those points which require further research.

In this thesis, an acoustic-phonetic approach to large-vocabulary Arabic word speech recognition has been demonstrated. In this approach a broad phonetic classification technique is used instead of detailed phonetic analysis to overcome the variabilities in the acoustic realisation of utterances. The broad phonetic description of a word is used as a means of lexical access, where the lexicon is structured into sets of words sharing the same broad phonetic labelling. The phonetic structure of the chosen set of words selects and schedules the context-dependent procedures which are most appropriate for performing detailed verification analyses in order to determine the most likely spoken one among the word candidates. Our aim is to try as much as possible to identify all words uniquely by broad phonetic representation without detailed acoustic analysis.

The work in this thesis can be divided into two main parts. The first part investigates the efficiency of the broad phonetic classification technique, and the effectiveness of different classification schemes on the number of words sharing the same phonetic labelling using a lexicon of 10,000 words [Chapter 4]. The second part concerns the implementation of a speech recognition system based on the proposed broad phonetic segmentation technique [Chapters 5,6 and 7]. Novel procedures have been developed in different parts of the speech recognition system. The performance of the implemented procedures has been evaluated by using 570 words which comprise the repetition of 100 Arabic words uttered by several speakers.

8.2 Phonetic Classification Schemes

Ten different phonetic classification schemes have been investigated using two lexicons. The first lexicon comprises the most frequent 3000 words in the Arabic language, and the second lexicon contains the first lexicon plus some of the derivatives of its words and other randomly chosen words. In each classification scheme, different sets of broad phonetic classes are employed. The ten classification schemes are divided into two similar groups. In one group, the six vowels (used in the Arabic language) are classified into one class, while in the other group the vowels are classified according to their phonemic forms. The 29 consonants (used in the Arabic language) in the two classification groups are classified into one, four, five, seven, or eleven classes. Thus, the broad phonetic description varies from a very rough one (only 2 classes, i.e., vowel and consonant) to one close to the phonemic form (17 classes).

The statistical results show that about 38% of the tested 10,000 words can be uniquely represented by using 8 broad phonetic classes (i.e., 7 classes for consonants and one class for vowels). In this case, the maximum set-size (i.e., the maximum number of words sharing the same labelling) in the lexicon was 28 words. When detailed vowel classification (according to their phonemic forms) is introduced yielding a total of 13 phonetic classes, the percentage of uniquely represented words rises to 83%. In this case, the maximum set-size was 6 words [Chapter 4]. An 88% of the words are uniquely represented when using 17 phonetic labels (6 for vowels and 11 for consonants), where the maximum set-size was 5 words. These results suggest that a fully detailed phonetic analysis of the speech signal is perhaps unnecessary.

The statistical studies also show the following results. The six vowels represent about 43% of the total number of phonemes, while the 29 consonants represent 57%. The vowels /a/ and /aa/ represent about 60% of the total number of vowels. With regards to the syllabic types used in the language, the results show that the syllable /consonant-vowel/ represents about 67% of the total number of syllables in the 10,000 words lexicon. This means, that the probability of having a consonant cluster comprising more than one consonant is about 0.33.

The statistical studies which are reported in this thesis regarding the discrimination of words in a large vocabulary lexicon according to different broad phonetic classification schemes are the first and only studies applied to the Arabic language so far.

8.3 The Word Recognition Model

In the adopted word recognition model [Chapter 4], the speech signal of a word is segmented according to one of the proposed phonetic classification schemes. In the implemented scheme, consonants are classified into four broad phonetic classes, i.e., voiced plosive, unvoiced plosive, unvoiced fricative, and other voiced consonants. The vowels are described by their phonemic form, i.e., 'a', 'aa', 'u', 'uu', 'i', and 'ii'.

An unknown word at the input of the recognition system is described by a string of broad phonetic labels. This has been achieved by five main procedures which are:

- Voiced-unvoiced-silence (V-UV-S) segmentation.
- Vowel detection and identification.
- Spectral transition detection.
- Initial segmentation and labelling.
- Segmentation error correction.

The V-UV-S segmentation procedure gives almost perfect results due to the post-processing (editing) performed on the initial results. In the post-processing, spurious segments are removed by a non-linear smoothing process.

In the vowel detection process, two concepts have been introduced, i.e., the vowel estimated duration and the vowel representative frame. Also, a heuristic method has been developed to distinguish between short and long vowels. The accuracy of the detection process is about 99%. Unfortunately, this 1% error rate leads to a higher percentage of errors at the end of the segmentation process. Thus, the accuracy of the vowel detection algorithm is very important. This 1% error is caused by the failure of the detection algorithm to detect those vowels which have relatively less total energy than some adjacent voiced consonants. However, further research is required to develop a special algorithm to detect such vowels. This algorithm may use the energy in selected frequency bands or other parameters.

The vowel identification process determines the vowel identity using the vowel representative frame. Two methods were applied, i.e., vector quantisation and formant methods. The accuracy of the vector quantisation method is 99%, while the formant method has given almost perfect accuracy. It is preferable to use the formant method

because of the following reasons:

- It gives extra information about vowels in pharyngealised context leading to extra cues for further consonant classification.
- It gives more freedom in a multi-speaker system after incorporating a formant normalisation algorithm.

Thus, in this respect two aspects require further work. The first one is the implementation of a formant normalisation algorithm. The second aspect is on improving the formant estimation algorithm to detect higher formants for the purpose of normalisation. The implemented formant estimation method is based on the phase spectrum of the LPC model. The choice of the LPC model's order determines the level and the quantity of the spectral details. To resolve the case of two closely spaced formants, the model's order should be high, but this will lead to the presence of spurious spectral peaks. These spurious peaks make the choice of the actual formants very difficult to achieve. Recently, a new approach has been proposed [143] to estimate the formant frequencies without modelling the speech signal, hence avoiding the above-mentioned problem related to the model's order. In this approach, formants are extracted from the phase spectrum of the speech signal after converting the speech signal (in the frame under analysis) to a minimum phase signal. The length of the analysis window for formant extraction is then chosen less than the pitch period in order to maintain a smooth frequency response devoid of fluctuation. Thus, this method can be incorporated in our system, and further work is needed to select the length of the analysis window which leads to better results.

The spectral transition detection procedure seems to work well [Chapter 6]. The results of the transition detection are reflected in the spectral variation contour of each word. This contour displays almost all the transitions which can be noticed in the speech spectrogram of a certain word. If a transition is not clear in the speech spectrogram of a word, it will not be reflected on the spectral variation contour of that word. The only problem in the transition detection is the case of having some missing peaks (transitional points) along the spectral variation contour due to the smoothing process which is applied to the spectral parameters. Although such peaks are recovered during the correction procedure, special attention should be given to such cases. Another point which requires some further study is the possibility of using a fixed threshold for all words to detect the prominent peaks (transitional points) of the spectral variation contour. The current threshold is taken as the mean value of the spectral variation function of each word.

Having a fixed threshold leads to faster processing where there will be no need to wait until the word ends to compute the mean value of the spectral variation function. Also, the availability of such a fixed threshold will facilitate the application of the transition detection algorithm to continuous speech.

The results of the V-UV-S segmentation, the vowel detection and identification, and the transition detection are organised into a special table called the segmentation result table [Chapter 7]. Differences of less than a certain limit between the transitional points of the V-UV-S contour and the spectral variation contour are tolerated. Subject to the content of the result table, the segments on the spectral variation contour are labelled according to the considered classification scheme.

Most of the errors which are caused by the previous procedures are detected and corrected during the error correcting procedure. This procedure employs durational information (such as the vowel estimated duration and the segment length) in addition to some phonological constraints. The correction procedure deals with all the detected and expected errors, and the segmentation result table is updated after each correction step. As for the unforeseen errors at this stage, a special correction algorithm can be designed to account for any newly discovered error and can easily be incorporated in the system. Such an algorithm uses the information in the result table at its input and updates this table after any modification.

The above five procedures are designed in such a way that modifying one procedure will not affect the others, provided that the required information is filled at the end in the result table of each word. For example, modifying the vowel detection and identification procedure to improve its performance will not affect the performance of other procedures, provided that the modified procedure will supply the result table of a certain word with the vowel estimated durations, the vowel representative frames and the vowel identities of that word.

8.4 Further Consonant Classification

The implemented labelling scheme uses four labels for consonant description. Further labelling is also possible where the class 'voiced consonant' comprises nasal, semivowel, liquid, and voiced fricative consonants. In addition, the plosive and fricative classes can be split into pharyngealised or plain (non-pharyngealised) classes [Chapters 3 and 4]. In

the implemented recognition system, the segment boundaries are already determined and given in the segmentation result table. Also, the segment identity (vowel or consonant) is known and given in the result table. Thus further consonant labelling can be achieved by introducing two special procedures. One of them can be called the voiced consonant procedure, and is used to split voiced consonant segments into different possible classes. The other one can be called the pharyngealised consonant procedure, and is used to decide whether a plosive or fricative consonant segment is pharyngealised or plain. It has been shown [Chapter 5] that the difference between the formant frequencies (F1 and F2) of a certain vowel can give an indication of the presence of a pharyngealised consonant in its carrier word.

The purpose of this further consonant classification is to reduce the number of word candidates which share the same phonetic labelling with the input or the test word. This will reduce the amount of detailed phonetic analysis at the verification stage of the recognition system [Chapter 4] to select the most likely word candidate. However, the system can start with performing vowel recognition and consonant classification according to the implemented scheme (4 labels for consonants). The resultant string of labels is used for lexical access. If the number of word candidates is more than one, the system calls the voiced consonant procedure. Then, the modified string of labels is reused for lexical access. If the number of word candidates is still more than one, the system calls the pharyngealised consonant procedure. In this case, if the number of word candidates exceeds one, the verification process is to be carried out. An alternative solution is to maintain the lexical access at the implemented classification scheme, and further consonant classification can be achieved directly in the verification stage of the recognition model only when it is required. In the latter case, the verifier may use mathematical approaches such as hidden Markov modelling or template matching for further consonant classification.

In this thesis, the physical correlates of the prosodic features, i.e., energy, pitch, and duration, have been measured and used in different parts of the speech recognition system. They can also be used to extract other features (such as stress). These features, besides other phonological constraints (such as consonant dissociation), can be used in the verification stage in order to avoid as much as possible the detailed consonant analysis. The extraction of such reliable features requires further research.

8.5 Closing Remarks

The accuracy of the segmentation and labelling process in the proposed speech recognition model is relatively high (95%). Improving the accuracy of the vowel recognition procedure will surely improve the accuracy of the whole system.

Those points which require further work are summarised as follows:

- Improving the performance of the vowel detection and vowel identification processes.
- Investigating the possibility of using a fixed threshold in the detection of the prominent peaks (the transitional points) on the spectral variation contour of all words.
- Developing a special procedure to split the voiced consonant segments into further detailed classes.
- Developing a special procedure to distinguish between pharyngealised and non-pharyngealised consonants.
- Developing the verification stage in the recognition model.

In addition to the above points, further work is needed to test the performance of the developed algorithms using speech degraded by noise, telephone bandwidth speech, and speech uttered by uncooperative speakers.

Most of the developed procedures for the word recognition system may be modified and used in the acoustic-phonetic processor of a continuous speech recognition system.

The developed recognition method can be implemented almost in real time (where some decisions are delayed until after the end of the test word), using the already available digital signal processing chips.

References

- [1] M. Mrayati and M. Al-Zabibi, "Contribution to Computer Arabization", Int. Symposium on Transfer of Computer Technology, Amman, JORDAN, Nov. 1987.
- [2] M. Mrayati, "Speech Processing Application to the Arabic Language", Workshop on Computer Processing of the Arabic language, KUWAIT, April 1985.
- [3] M. Mrayati and J. Makhoul, "Man-Machine Communication and the Arabic Language", Proc. Arab School on Science and Technology: "Applied Arabic Linguistics and Signal & Information Processing", Rabat, MOROCCO, Sept. 1983.
- [4] A. Mouradi, A. Rajouani, and M. Najim, "Unlimited Vocabulary Synthesis System for Arabic Language", Proc. 4th Int. Conf. Digital Processing Signal in Communications, pp. 329-331, U.K., April 1985.
- [5] Y. A. EL-Imam, "Speech Synthesis by Concatenating Sub-Syllabic Sound Units", Proc. IEEE ICASSP-87, pp. 2416-17, April 1987.
- [6] M. Mrayati, R. Carre, and B. Guerin, "Distinctive Regions and Modes: A New Theory of Speech Production", Speech Communication, vol. 7, pp. 257-286, Oct. 1988.
- [7] M. A. Hashish, A. T. El-Kheshen, and M. R. El-Ghonemy, "Experience in Isolated Arabic Word Recognition", Workshop on Computer Processing of the Arabic language, KUWAIT, April 1985.
- [8] O. S. Emam and M. A. Hashish, "Recognition of Isolated Arabic Digits using Hidden Markov Models", Proc. European Conf. on Speech Technology, pp. 312-316, U.K., Sept. 1987.

- [9] R. L. Brewster, A. Al-Otaibi, and Y. El-Imam, "Automatic Arabic Vowel Recognition", Proc. IERE conf. Digital Communication, pp. 303-306, U.K., Sept. 1988.
- [10] D. A. Abduh, "The Common Words in the Arabic Language", Publication of Riyadh University, SAUDI ARABIA, 1979, (in Arabic).
- [11] J. Flanagan, "Voice of Man and Machines", J. Acoust. Soc. Am., vol. 51, pp. 1375-1387, May 1972.
- [12] H. Dudley, "Remaking Speech", J. Acoust. Soc. Am., vol. 11, pp. 169-177, 1939.
- [13] R. Potter, G. Kopp, and H. G. Green, Visible Speech, Van Nostrad, New York, 1947.
- [14] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic Recognition of Spoken Digits", J. Acoust. Soc. Am., vol. 24, pp. 637-642, Nov. 1952.
- [15] H. Dudley and S. Balashek, "Automatic Recognition of Phonetic Patterns in Speech", J. Acoust. Soc. Am., vol. 30, pp. 721-732, Aug. 1958.
- [16] H. F. Olson and H. Belar, "Phonetic Typewriter", J. Acoust. Soc. Am., vol. 28, pp. 1072-1081, Nov. 1956.
- [17] J. Wiren and H. L. Stubbs, "Electronic Binary Selection System for Phoneme Classification", J. Acoust. Soc. Am., vol. 28, pp. 1082-1091, Nov. 1956.
- [18] J. W. Forgie and C. D. Forgie, "Results Obtained from a Vowel Recognition Computer Program", J. Acoust. Soc. Am., vol. 31, pp. 1480-1489, Nov. 1959.
- [19] P. Denes and M. V. Mathews, "Spoken Digit Recognition using Time-Frequency Pattern Matching", J. Acoust. Soc. Am., vol. 32, pp. 1450-1455, Nov. 1960.

- [20] T. Sakai and S. Doshita, "The Automatic Recognition System for Conversational Sound", *IEEE Trans. Electronic Computers*, vol. EC-12, pp. 836-846, Dec. 1963.
- [21] J. W. Cooley and J. W. Tukey, "An Algorithm for the Machine Calculation of Complex Fourier Series", *Mathematics of Computation*, vol. 19, pp. 297-301, 1965.
- [22] D. H. Klatt, "Review of the ARPA Speech Understanding Project", *J. Acoust. Soc. Am.*, vol. 62, pp. 1345-1366, Dec. 1977.
- [23] B. T. Lowerre and D. R. Reddy, "The Harpy Speech Understanding System", in *Trends in Speech Recognition*, Ed. W. A. Lea, Prentice-Hall, 1980.
- [24] D. R. Reddy, D. L. Erman, and R. B. Neely, "A Model and a System for machine recognition of Speech", *IEEE Trans. Audio and Electroacoustic*, vol. AU-21, pp. 229-238, June 1973.
- [25] J. J. Wolf and W. A. Woods, "The HWIM Speech Understanding System", in *Trends in Speech Recognition*, Ed. W. A. Lea, Prentice-Hall, 1980.
- [26] J. K. Baker, "The Dragon System, an Overview", *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-23, pp. 24-29, Feb. 1975.
- [27] F. Jelinek, L. R. Bahl, and R. L. Mercer, "Design of Linguistic Statistical Decoder for the Recognition of Continuous Speech", *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 250-256, May 1975.
- [28] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, Springer-Verlag 1965.
- [29] J. Makhoul, "Linear Prediction: A Tutorial Review", *Proc. IEEE*, vol. 63, pp. 561-580, April 1975.
- [30] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

- [31] G. White and R. B. Neely, "Speech Recognition Experiments with Linear Prediction, Bandpass Filtering, and Dynamic Programming", *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-27, pp. 183-188, April 1979.
- [32] B. A. Doutrich, L. R. Rabiner, and T. Martin, "On the Effect of Varying Filter Bank Parameters on Isolated Word Recognition", *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-31, pp. 793-806, Aug. 1983.
- [33] A. V. Oppenheim and R. W. Schaffer, *Digital Signal Processing*, Prentice-Hall, 1975.
- [34] L. R. Rabiner and R. W. Schaffer, *Digital Processing of Speech Signal*, Prentice-Hall, 1978.
- [35] S. B. Davis and P. Mermelestein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-28, pp. 357-366, Aug. 1980.
- [36] E. Zwicker, "Subdivision of the Audible Frequency Range into Critical Bands", *J. Acoust. Soc. Am.*, vol. 33, p. 248, Feb. 1961.
- [37] E. Zwicker and E. Terhardt, "Analytical Expression for Critical-Band Rate and Critical Bandwidth as a Function of Frequency", *J. Acoust. Soc. Am.*, vol. 68, pp. 1523-1525, Nov. 1980.
- [38] K. Shirai and M. Honda, "Estimation of Articulatory Parameters from Speech Waves", *Electronic and Communication in Japan*, vol. 61, No. 5, pp. 1-8, May 1978.
- [39] R. Carre and M. Mrayati, "New Concept in Acoustic-Articulatory-Phonetic relation Perspectives and Application", *Proc. IEEE ICASSP-89*, pp. 231-234, May 1989.
- [40] M. Blomberg, R. Carlson, K. Elenius, and B. Granstrom, "Auditory Models in Isolated Word Recognition", *Proc. IEEE ICASSP-84*, pp. 17.9.1-4, 1984.

- [41] E. Zwicker and E. Terhardt, "Automatic Speech Recognition Using Psychoacoustic Models", *J. Acoust. Soc. Am.*, vol. 65, pp.478-489, Feb. 1979.
- [42] J. Cohen, "Application of an Adaptive Auditory Model to Speech Recognition", *J. Acoust. Soc. Am.*, vol. 78, p. s50(A), Feb. 1985.
- [43] L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances", *B. Sys. Tech. J.*, vol. 54, pp. 297-315, Feb 1975.
- [44] L. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "An improved Endpoints Detector for Isolated Word Recognition", *IEEE Trans. Speech Acoust. Signal Processing*, vol. ASSP-29, pp. 777-785, Aug. 1981.
- [45] H. Ney, "An optimization Algorithm for Determining the Endpoints of Isolated Utterances", *Proc. IEEE ICASSP-81*, pp. 720-723, 1981.
- [46] J. G. Wilpon, L. R. Rabiner, and T. Martin, "An improved Word-Detection Algorithm for Telephone-Quality Speech Incorporating Both Syntactic and Semantic Constraints", *AT&T B. Labs. Tech. J.*, vol. 63, pp. 479-497, March 1984.
- [47] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition", *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-23, pp. 67-72, Feb. 1975.
- [48] H. Sakoe, S. Chiba, "Dynamic Programing Algorithm Optimization for Spoken Word Recognition", *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-26, pp. 43-49, Feb. 1978.
- [49] K. K. Paliwal, A. Agarwal, and S. Sinha, "A Modification over Sakoe and Chiba's Dynamic Time Warping Algorithm for Isolated Word Recognition", *Proc. IEEE ICASSP-82*, pp. 1259-1261, May 1982.

- [50] L. R. Rabiner, A. Rosenberg, and S. Levinson, "Consideration in Dynamic Time Warping Algorithm for Discrete Word Recognition", IEEE Trans. Speech Acoust. Signal Processing, vol. ASSP-26, pp. 575-582, Dec 1978.
- [51] C. Myers, L. R. Rabiner, and A. Rosenberg, "Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated word Recognition", IEEE Trans. Acoust. Speech Signal Processing, vol. ASSP-28, pp. 623-635, Dec. 1980.
- [52] S. Haltsonen, "Improved Dynamic Time Warping Methods for Discrete Utterance Recognition", IEEE Trans. Acoust. Speech Signal Processing, vol. ASSP-33, pp. 449-450, April 1985.
- [53] H. Hohne, C. Coher, S. Levinson, and L. R. Raabiner, "On Temporal Alignment of Sentences of Natural and Synthetic Speech", IEEE Trans. Speech Acoust. Signal Processing, vol. ASSP-31, pp. 807-813, Aug. 1983.
- [54] H. Leung and V. W. Zue, "A Procedure for Automatic Alignment of Phonetic Transcriptions With Continuous Speech," Proc. IEEE ICASSP-84, pp. 2.7.1-4, 1984.
- [55] A. H. Gray and J. D. Markel, "Distance measures for Speech Processing", IEEE Trans. Acoust. Speech Signal Processing, vol. ASSP-24, pp. 380-391, Oct. 1976.
- [56] R. Pieranccini, "Pattern Compression in Isolated Word Recognitions", Signal Processing, vol. 7, pp. 1-15, Sept. 1984.
- [57] V. Gupta, M. Lenning, and P. Mermelstein, "Decision Rules for Speaker-Independent Isolated Word Recognition", Proc. IEEE ICASSP-84, pp. 9.2.1.-4, 1984.
- [58] L. R. Rabiner, S. Levinson, A. E. Rosenberg, and J. Wilpon, " Speaker-Independent Recognition of Isolated Words Using Clustering Techniques", IEEE Trans. Acoust. Speech Signal Processing, vol. ASSP-27, pp. 336-349, Aug. 1979.

- [59] L. R. Rabiner and J. G. Wilpon, "A Simplified Robust Training Procedure for Speaker-Trained Isolated Word Recognition System", *J. Acoust. Soc. Am.*, vol. 68, pp. 1271-1276, Nov. 1980.
- [60] Y. J. Liu, "On Creating Averaging Templates", *Proc. IEEE ICASSP-84*, pp. 9.1.1- 4, 1984.
- [61] S. Levinson, L. R. Rabiner, A. Rosenberg, and J. Wilpon, "Interactive Clustering Techniques for Selecting Speaker-Independent Reference Templates for Isolated Word Recognition", *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-27, pp. 134-141, April 1979.
- [62] L. R. Rabiner and J. Wilpon, "Considerations in Applying Clustering Technique to Speaker-Independent Word Recognition", *J. Acoust. Soc. Am.*, vol. 66, pp. 663-673, Sept. 1979.
- [63] T. B. Martin, "Practical Applications of Voice Input to Machines", *Proc. IEEE*, vol. 64, pp. 487-501, Apr. 1976.
- [64] L. R. Rabiner and S. F. Levinson, "Isolated and Connected Word Recognition Theory and Selected Application", *IEEE Trans. comm.*, vol. COM-29, pp. 621-659., May 1981.
- [65] L. R. Rabiner and J. G. Wilpon, "Isolated Word Recognition Using a Two-Pass Pattern Recognition Approach", *Proc. IEEE ICASSP-81*, pp. 724-727, 1981.
- [66] T. Kaneko and N. R. Dixon, "A Hierarchical Decision Approach to Large-Vocabulary Discrete Utterance Recognition", *IEEE Trans. Speech Acoust. Signal Processing*, vol. ASSP-31, pp. 1061-1072, Oct. 1983.
- [67] M. Watari, M. Akabane, and Y. Sako, "A Speaker-Independent Word Recognition Based on Transient Matching", *Proc. IEEE ICASSP-83*, pp. 715-718, 1983.

- [68] L.R. Rabiner, K. Pan, and F. Soong, "On The Performance of Isolated Word Recognition Using Vector Quantization and Temporal Energy contours", AT&T. B. Labs. Tch. J., vol. 63, pp. 1245-1260, Sept. 1984.
- [69] J. E. Shore, D. K. Burton, "Discrete Utterance Speech Recognition Without Time Alignment", IEEE Tras. Inf. Theory, vol. IT-29, pp. 473-490, July 1983.
- [70] D. Burton, J. Shore, and J. Buck, " A Generalization of Isolated Word Recognition Using Vector Quantization", Proc. IEEE ICASSP-83, pp. 1021-1024, 1983.
- [71] D. Burton, J. Shore, and J. Buck, "Isolated Word Speech Recognition Using Multi-Section Vector Quantization Codebooks", IEEE Trans. Acoust. Speech Signal Processing, vol. ASSP-33, pp. 837-849, Aug. 1985.
- [72] L. E. Baum, " An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of a Markov Process", Inequalities, vol. 3, pp. 1-8, 1972.
- [73] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent Isolated Word Recognition", B. Sys.Tech. J., Vol. 62, pp. 1075-1105, April 1983.
- [74] G. D. Forney, "The Viterbi Algorithm", Proc. IEEE, vol. 61, pp. 268-278, March 1973.
- [75] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of Markov Process to Automatic Speech Recognition", B. Sys. Tech. J., vol. 62, pp. 1035-1074, April 1983.
- [76] L. R. Rabiner and B. H. Juang, "An Inroduction to Hidden Markov Models," IEEE ASSP Magazine, pp. 4-16, Jan. 1986.

- [77] B. H. Juang, "On the Hidden Markov Model and Dynamic Time Warping for Speech Recognition- a Unified View", AT&T B. Labs. Tech. J., vol. 63, pp. 1213-1243, Sept. 1984.
- [78] P. Lippmann, "An Introduction to Computing with Neural Nets", IEEE ASSP Magazine, pp. 4-22, April 1987.
- [79] A. Kranse and H. Hackbarth, "Scaly Artificial Neural Networks for Speaker-Independent Recognition of Isolated Words", IEEE Proc. ICASSP-89, pp. 21-24, 1989.
- [80] H. Sakoe, R. Isotani, K. Yoshida, K. Iso, and T. Watanabe, "Speaker-Independent Word Recognition Using Dynamic Time Programming Neural Networks", IEEE Proc. ICASSP-89, pp. 29-32, 1989.
- [81] D. H. Klatt, "Scriber and LAFS: Two New Approaches to Speech Analysis", in Trends in Speech Recognition, Ed., W. A. Lea, Prentice-Hall, 1980.
- [82] W. A. Ainsworth, "Speech Recognition by Machine", Peter Peregrinus Ltd., 1988.
- [83] O. Fujimura, "Syllable as unit of Speech Recognition", IEEE Trans. Speech Acoust. Signal Processing, vol. ASSP-23, pp. 82-86, Feb. 1975.
- [84] G. Ruske, "Auditory Perception and its Application to Computer Analysis of Speech", in Computer Analysis and perception vol. II, Auditory Signal, Eds. C. Y. Suen, and R. De Mori, CRC Press Inc., 1982.
- [85] J. G. Wilpon, B. H. Juang, and L. R. Rabiner, "An Investigation on the Use of Acoustic-sub-word Units for Automatic Speech Recognition", Proc. IEEE ICASSP-87, pp. 20.7.1- 4, 1987
- [86] C. H. Lee, B. H. Juang, F. K. Soong, and L. R. Rabiner, "Word Recognition Using Whole Word and Sub-word Models", Proc. IEEE ICASSP-89, pp. 683-686, 1989.

- [87] P. B. Denes and T. G. Von Keller, "Articulatory Segmentation for Automatic Recognition of Speech", The 6th Int. Congress on Acoustic, Tokyo, JAPAN, pp. B. 143-146, Aug. 1968.
- [88] R. De Mori, Computer Models of Speech Using Fuzzy Algorithm, Pleum Press, 1983.
- [89] E. C. Bronson, "Syntactic Pattern Recognition of Discrete Utterances", Proc. IEEE ICASSP-83, pp. 719-722, 1983.
- [90] M. R. Sambur and L. R. Rabiner, "A statistical Decision Approach to the Recognition of Connected Digits", IEEE Trans. Acoust. Speech Signal Processing, vol. ASSP-24, pp. 550-558, 1976.
- [91] C. S. Myers and L. R. Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition", IEEE Trans. Acoust. Speech Signal Processing, vol. ASSP-29, pp. 284-296, 1981.
- [92] H. Ney, "The Use of One-Stage Dynamic Programming Algorithm for Connected Word Recognition", IEEE Trans. Acoust. Speech Signal Processing, vol. ASSP-32, pp. 263-271, April 1984.
- [93] R. A. Cole, A. I. Rudnicky, V. W. Zue, and D. R. Reddy, "Speech as Patterns on Paper", in Perception and Production of Fluent Speech, Ed. R. A. Cole, Lawrence Erlbaum Associates, pp. 3-50, 1980.
- [94] J. E. Shoup, "Phonological Aspects of Speech Recognition", in Trends in Speech Recognition, Ed. W. A. Lea, pp. 125-138, Speech Science Publication, 1980.
- [95] W. A. Lea, "Prosodic Aids to Speech Recognition", in Trends in Speech Recognition, Ed. W. A. Lea, pp. 166-205, Prentice-Hall, 1980.
- [96] W. A. Ainsworth, Mechanisms of Speech Recognition, Pergamon Press, 1976.

- [97] L. D. Erman and V. R. Lesser, "The Hearsay II Speech Understanding System", in Trends in Speech Recognition, Ed. W. A. Lea, prentice-Hall, 1980.
- [98] J. S. Bridle, "An Efficient Elastic-Template Method for Detecting Given Words in Running Speech", British Acoustical Society Meeting, pp. 1-4, April 1973.
- [99] R. W. Christiansen and C. K. Rushforth, " Detecting and Locating Key Words in Continuous Speech Using Linear Predictive Coding", IEEE Trans. Acoust. Speech Signal Processing, vol. ASSP-25, pp. 361-367, Oct. 1977.
- [100] R. E. Nolford, " The Enhancement of Word Spotting Techniques", Proc. IEEE ICASSP-80, pp. 209-212, 1980.
- [101] J. K. Baker, "Stochastic Modelling for Automatic Speech Understanding", in Speech Recognition, Ed. D. R. Reddy, pp. 521-542, Academic Press 1975.
- [102] G. Fant, Speech Sounds and Features, The MIT Press, 1973.
- [103] S. H. Alani, "Arabic Phonology: An Acoustic and Physiological Investigation", Ph.D. Thesis, Indiana University, USA 1966.
- [104] G. Ghazali, "Element of Arabic Phonetics", Proc. Arab School on Science and Technology, "Applied Arabic Linguistics and Signal & Information Processing", Rabat, MOROCCO, Sept. 1983.
- [105] G. K. Pullum and W. A. Ladusaw, Phonetic Symbol Guide, The University of Chicago Press, 1986.
- [106] M. Al-Zabibi and S. Datta, "Arabic Vowel Recognition in a Large Vocabulary Isolated Word Speech Recognition System", Proc. Gulf Digital Signal Processing Symp., KUWAIT, May 1990.
- [107] A. H. Moussa, "Statistical Study of Arabic Roots in Moijam Lisan Al-Arab", Kuwait University Press, 1972, (in Arabic).

- [108] M. Mrayati, "Statistical Studies of Arabic language Roots", Proc. Arab School of Science and Technology, "Applied Arabic Linguistics and Signal & Information Processing", Rabat, MOROCCO, Sept. 1983.
- [109] A. H. Moussa, "Occurrence Probability of Characters and Entropy of Arabic Language", Proc. Arab School of Science and Technology, "Informatics and Applied Arabic Linguistics", Zabadani, Syria, Aug. 1985.
- [110] M. Bawab, "Statistical Studies of Arabic Language", unpublished Study, Scientific Studies and Research Centre, Damascus, Syria, 1985.
- [111] A. S. Shaheen, "Phonological Method for Arabic Structure", Alrisalah Press, 1985, (in Arabic).
- [112] Y. Meer-Alam and H. Tayyan, "Arabic Lexicon: Statistical and Phonological Study on Arabic Language", MA Thesis, Damascus University, SYRIA, 1984, (in Arabic).
- [113] H. Tayyan, Y. Meer-Alam, and M. Mrayati, "Data Base for Arabic Roots", 2nd. Conf. on Arabic Computational Linguistics, KUWAIT, Nov. 1989, (in Arabic).
- [114] Y. Haydar and M. Mrayati, "Study on the Intonation of Arabic Standard Phrases", Strasbourg Institute of Phonetic Report, Strasbourg, France, 1986, (in French).
- [115] D. W. Shipman and V. W. Zue, "Properties of Large Lexicons: Implication for Advanced Isolated Word Recognition System", Proc. IEEE ICASSP-82, pp. 546-549, 1982.
- [116] A. Waibel, "Suprasegmentals in Very Large Vocabulary Isolated Word Recognition", Proc. IEEE ICASSP-84, pp. 26.3.1-4, 1984.
- [117] D. P. Huttenlocher and V. W. Zue, "A Model of Lexical Access from Partial Phonetic Information", Proc. IEEE ICASSP-84, pp. 26.4.1- 4, 1984.

- [118] G. Vernooij, G. Bloothooff, and Y. V. Holsteijn, "A Simulation Study on the Usefulness of Broad Phonetic Classification in Automatic Speech recognition", Proc. IEEE ICASSP-89, pp. 85-88, 1989.
- [119] A. Giordana and L. Saitta, "Discrimination of Words in a Large Vocabulary Using Phonetic Description", Int. J. Man-Machine Studies, vol. 24, pp. 453-473, 1986.
- [120] Y. Hlal, "Morphological Analysis of Arabic", Proc. Arab School of Science and Technology, "Informatics and Applied Arabic Linguistics", Zabadani, Syria, Jul. 1985.
- [121] M. Bawab, Y. Meer-Alam, H. Tayyan, and M. Mrayati, "Computerized Arabic Morphological Analyzer-Generator", Proc. 4th Int. Meeting on Linguistics, TUNISIA, Nov. 1987, (in Arabic).
- [122] N. H. Hegazi and A. A. Elsharkawi, "An Approach to a Computerised Lexical Analyzer for Natural Arabic Text", Workshop on Computer Processing of the Arabic language, KUWAIT, April 1985.
- [123] S. S. Al-Fadaghi and H. B. Al-Sadoun, "Morphological Compression of Arabic Text", Electrical and Computer Engineering Dept. Report, Kuwait University, KUWAIT 1989.
- [124] S. Datta, and M. Al-Zabibi, "Discrimination of Words in a Large Vocabulary Speech Recognition System", Proc. Int. Conf. Spoken Language Processing, ICSLP-90, Kobe JAPAN, Nov. 1990.
- [125] B. S. Atal and L. R. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Application to Speech Recognition", IEEE Trans. Acoust. Speech Signal Processing, vol. ASSP-24, pp. 201-212, June 1976.

- [126] R. L. Rabiner, C. E. Schmidt, and B. S. Atal, "Evaluation of a Statistical Approach to Voiced-Unvoiced-Silence Analysis for Telephone-Quality Speech", *B. Sys. Tech. J.*, vol. 56, pp. 455-481, March 1977.
- [127] S. Knorr, "Reliable Voiced/Unvoiced Decision", *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-27, pp. 263-267, June 1979.
- [128] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms", *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-24, pp. 399-417, Oct. 1976.
- [129] M. M. Sondhi, "New Methods of Pitch Extraction", *IEEE Trans. Audio Electroacoustic*, vol. AU-16, pp. 262-266, June 1968.
- [130] A. M. Noll, "Cepstral Pitch Determination", *J. Acoust. Soc. Am.*, vol. 41, pp. 293-309, Feb. 1967.
- [131] R. W. Schafer, L. R. Rabiner, and O. Herrmann, "FIR Digital Filter Banks for Speech Analysis", *B. Sys. Tech. J.*, vol. 54, pp. 531-544, March 1975.
- [132] L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing", *IEEE Trans. Speech Acoust. Signal Processing*, vol. ASSP-23, pp. 552-557, Dec. 1975.
- [133] L. R. Rabiner and F. K. Soong, "Single Frame Vowel Recognition Using Vector Quantization with Several Distance Measures", *AT&T Tech. J.*, vol. 64, pp. 2319-2330, Dec. 1985.
- [134] S. S. Stevens, "The Measurement of Loudness", *J. Acoust. Soc. Am.*, vol. 27, pp. 815-829, Sept. 1955.
- [135] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantization", *IEEE Trans. Commun.*, vol. COM-28, pp. 84-95, Jan. 1980.

- [136] A. Buzo, A. H. Gray, R. M. Gray, and J. D. Markel, "Speech Coding Based Upon Vector Quantization", IEEE Trans. Acoust. Speech Signal Processing, vol. ASSP-28, pp. 562-574, Oct. 1980.
- [137] L. R. Rabiner, M. M. Sondhi, and S. E. Levinson, " Note on the Properties of a Vector Quantizer for LPC Coefficients", B. Sys. Tech. J. , vol. 62, pp.2603-2616, Oct. 1983.
- [138] H. Wakita, "Normalization of Vowels by Vocal-Tract Length and its Application to Vowel Identification", IEEE Trans. Acoust. Speech Signal Processing, vol. ASSP-25, pp. 183-192, April 1977.
- [139] B. Yegnanarayana, "Formant Extraction from Linear Prediction Phase Spectra", J. Acoust. Soc. Am., vol.63, pp.1638-1640, May 1978.
- [140] G. F. Chollel and C. Gagnoulet, " On the Evaluation of Speech Recognizers and Databases Using A Reference System", Proc. IEEE ICASSP-82, pp. 2026-2029, 1982.
- [141] C. Gagnoulet and M. Couvrat, "SERAPHINE: A Connected Word Speech Recognition System", Proc. IEEE ICASSP-82, pp. 887-890 , 1982.
- [142] H. Hermansky, B. A. Hanson, and H. Wakita, " Low Dimensional Representation of Vowels Based on All-Pole Modeling in the Psychophysical Domain", Speech Communication, vol. 4, pp. 181-187, Aug. 1985.
- [143] G. Duncan, B. Yegnanarayana, and H. A. Murthy, "A Nonparametric Method of Formant Estimation Using Group Delay Spectra", Proc. IEEE ICASSP-89, pp. 572-575 , 1982.

Appendix A

A List of the words in the speech database

In this appendix a list of the 100 words which have been used in the speech database. The list shows the following for each word:

- The phonemic description
- The syllabic pattern
- The phonetic description according to the implemented classification scheme (i.e., four labels 'VP', 'UP', 'UF', and 'VC' are used for consonants, and six labels according to their phonemic form are used for vowels).
- The English equivalent or meaning
- The Arabic script

We should mention here, that each word may have different meaning according to the context. In this list, we have chosen the most common meaning of each word.

| | | | | |
|--------|---------|---------------|--------------------|------|
| fii | CVV | UF-ii | 'in' | في |
| laa | CVV | VC-aa | 'no' | لا |
| min | CVC | VC-i-VC | 'from' | من |
| sifr | CVCC | UF-i-UF-VC | 'zero' | صفر |
| mæa | CV-CV | VC-a-VC-a | 'with' | مع |
| ɛalaa | CV-CV | VC-a-VC-aa | 'on' | على |
| ʔilaa | CV-CVV | VC-i-VC-aa | 'from' | الى |
| kalam | CV-CVC | UP-a-VC-a-VC | 'pen' | قلم |
| ʃamaal | CV-CVVC | UF-a-VC-aa-VC | 'north' | شمال |
| ʒanuub | CV-CVVC | UF-a-VC-uu-VP | 'south' | جنوب |
| niʒaam | CV-CVVC | VC-a-VC-aa-VC | 'system' | نظام |
| kabiir | CV-CVVC | UP-a-VP-ii-VC | 'big' or 'large' | كبير |
| ṣayir | CV-CVVC | UF-a-VC-ii-VC | 'small' | صغير |
| ʃariif | CV-CVVC | UF-a-VC-ii-UF | 'honest' | شريف |
| wisaam | CV-CVVC | VC-i-UF-aa-VC | (name of a male) | وسام |
| wiṣaal | CV-CVVC | VC-i-UF-aa-VC | (name of a female) | وصال |
| kaana | CVV-CV | UP-aa-VC-a | 'was' | كان |
| kaala | CVV-CV | UP-aa-VC-a | 'he said' | قال |
| naama | CVV-CV | VC-aa-VC-a | 'he slept' | نام |
| faaza | CVV-CV | UF-aa-VC-a | 'he won' | فاز |
| saaka | CVV-CV | UF-aa-UP-a | 'he drove' | ساق |
| waahid | CVV-CVC | VC-aa-UF-i-VP | 'one' | واحد |

| | | | | |
|---------|----------|------------------|----------------------|-------|
| naakis | CVV-CVC | VC-aa-UP-i-UF | 'minus' | ناقص |
| zaaʔid | CVV-CVC | VC-aa-UP-i-VP | 'plus' | زائد |
| haamiʃ | CVV-CVC | UF-aa-VC-i-UF | 'margin' | هامش |
| xamsa | CVC-CV | UF-a-VC-UF-a | 'five' | خمسة |
| sitta | CVC-CV | UF-i-UP-UP-a | 'six' | سته |
| sabes | CVC-CV | UF-a-VP-VC-a | 'seven' | سبعه |
| tisea | CVC-CV | UP-i-UF-VC-a | 'nine' | تسعه |
| maʃfaa | CVC-CVV | VC-a-UF-UF-aa | 'hospital' | مشفى |
| ʔawsaa | CVC-CVV | UP-a-VC-UF-aa | 'he advised' | أوصى |
| maysal | CVC-CVC | VC-a-VC-UF-a-VC | 'wash room' | مغسل |
| ʔakbar | CVC-CVC | UP-a-UP-VP-a-VC | 'bigger' or 'larger' | أكبر |
| ʔasʔar | CVC-CVC | UP-a-UF-VC-a-VC | 'smaller' | أصغر |
| ʔahsan | CVC-CVC | UP-a-UF-UF-a-VC | 'better' | أحسن |
| ʔafdal | CVC-CVC | UP-a-UF-VP-a-VC | 'best' | أفضل |
| ʔaswad | CVC-CVC | UP-a-UF-VC-a-VP | 'black' | أسود |
| ʔabjad | CVC-CVC | UP-a-VP-VC-a-VP | 'white' | أبيض |
| jaḥsub | CVC-CVC | VC-a-UF-UF-u-VP | 'he calculates' | يحسب |
| ʔiḥnaan | CVC-CVVC | UP-i-UF-VC-aa-VC | 'two' | إثنان |
| taksiim | CVC-CVVC | UP-a-UP-UF-ii-VC | 'division' | تقسيم |
| marwaan | CVC-CVVC | VC-a-VC-VC-aa-VC | (name of a male) | مروان |
| ʔassaan | CVC-CVVC | VC-a-UF-UF-aa-VC | (name of a male) | غسان |
| faʕala | CV-CV-CV | UF-a-VC-a-VC-a | 'he did' | فعل |

| | | | | |
|---------|-----------|-----------------|-------------------------|-------|
| masāḥa | CV-CV-CV | VC-a-UF-a-UF-a | 'he wiped' | مسح |
| ḏāḥaba | CV-CV-CV | UF-a-UF-a-VP-a | 'he went' | ذهب |
| kataba | CV-CV-CV | UP-a-UP-a-VP-a | 'he wrote' | كتب |
| ḡasala | CV-CV-CV | VC-a-UF-a-VC-a | 'he washed' | غسل |
| karaʔa | CV-CV-CV | UP-a-VC-a-UP-a | 'he read' | قرأ |
| ṭalaba | CV-CV-CV | UP-a-VC-a-VP-a | 'he requested' | طلب |
| naḏāra | CV-CV-CV | VC-a-VC-a-VC-a | 'he looked' | نظر |
| ḏaraba | CV-CV-CV | VC-a-VC-a-VP-a | 'he hit' or 'he strike' | ضرب |
| ʔakala | CV-CV-CV | UP-a-UP-a-VC-a | 'he ate' | أكل |
| ḥalaka | CV-CV-CV | UF-a-VC-a-UP-a | 'he shaved' | حلق |
| waḥaba | CV-CV-CV | VC-a-UF-a-VP-a | 'he jumped' | وثب |
| fataḥa | CV-CV-CV | UF-a-UP-a-UF-a | 'he opened' | فتح |
| wadaʕa | CV-CV-CV | VC-a-VP-a-VC-a | 'he put' | وضع |
| saʔala | CV-CV-CV | UF-a-UP-a-VC-a | 'he asked' | سأل |
| ṣafaha | CV-CV-CV | UF-a-UF-a-VC-a | 'he forgave' | صفح |
| ḥamala | CV-CV-CV | UF-a-VC-a-VC-a | 'he carried' | حمل |
| ḥasaba | CV-CV-CV | UF-a-UF-a-VP-a | 'he calculated' | حسب |
| ḥalaaḥa | CV-CVV-CV | UF-a-VC-aa-UF-a | 'three' | ثلاثة |
| jakuulu | CV-CVV-CV | VC-a-UP-uu-VC-u | 'he says' | يقول |
| janaamu | CV-CVV-CV | VC-a-VC-aa-VC-u | 'he sleeps' | ينام |
| takuulu | CV-CVV-CV | UP-a-UP-uu-VC-u | 'she says' | تقول |
| ʔaraada | CV-CVV-CV | UP-a-VC-aa-VP-a | 'he wanted' | أراد |

| | | | | |
|------------|---------------|-----------------------|---------------------|--------|
| ʔakaama | CV-CVV-CV | UP-a-UP-aa-VC-a | 'he established' | أقام |
| jusaawii | CV-CVV-CVV | VC-u-UF-aa-VC-ii | 'equal' | يساوي |
| nuuthihaa | CVV-CVV-CVV | VC-uu-UF-ii-VC-ii | 'we reveal it' | نوحىها |
| laakinna | CVV-CVC-CV | VC-aa-UP-i-VC-VC-a | 'but' | لكن |
| ʔarbaea | CVC-CV-CV | UP-a-VP-a-VC-a | 'four' | أربعة |
| markazu | CVC-CV-CV | VC-a-VC-UP-a-VC-u | 'centre' | مركز |
| jafealu | CVC-CV-CV | VC-a-UF-VC-a-VC-u | 'he does' | يفعل |
| jamsahu | CVC-CV-CV | VC-a-VC-UF-a-UF-u | 'he wipes' | يمسح |
| tamsahu | CVC-CV-CV | UP-a-VC-UF-a-UF-u | 'she wipes' | تمسح |
| tafealu | CVC-CV-CV | UP-a-UF-VC-a-VC-u | 'she does' | تفعل |
| jaʔkulu | CVC-CV-CV | VC-a-UP-UP-u-VC-u | 'he eats' | ياكل |
| juʔminu | CVC-CV-CV | VC-a-VC-VC-i-VC-u | 'he believes' | يؤمن |
| jahmilu | CVC-CV-CV | VC-a-UF-VC-i-VC-u | 'he carries' | يحمل |
| jaʔlubu | CVC-CV-CV | VC-a-UP-VC-u-VP-u | 'he requested' | يطلب |
| jaxtafii | CVC-CV-CVV | VC-a-UF-UP-a-UF-ii | 'he disappeared' | يختفي |
| ʔihtawaa | CVC-CV-CVV | UP-i-UF-UP-a-VC-aa | 'it contained' | احتوى |
| ʔassaala | CVC-CVV-CV | VC-a-UF-UF-aa-VC-a | 'washing machine' | غسالة |
| massaaha | CVC-CVV-CV | VC-a-UF-UF-aa-UF-a | 'wiper' or 'duster' | مساحه |
| θamaanija | CV-CVV-CV-CV | UF-a-VC-aa-VC-i-VC-a | 'eight' | ثمانيه |
| mataaliba | CV-CVV-CV-CV | VC-a-UP-aa-VC-i-VP-a | 'demands' | مطالب |
| kitaabunaa | CV-CVV-CV-CVV | UP-i-UP-aa-VP-uu-VC-a | 'our book' | كتابنا |
| ʔaasibunaa | CVV-CV-CV-CVV | UF-aa-UF-i-VP-u-VC-aa | 'our computer' | حاسبنا |

| | | | | |
|--------------|-------------------|--------------------------------|--|----------|
| jafəaluuna | CVC-CV-CVV-CV | VC-a-UF-VC-a-VC-uu-VC-a | 'they do' (plural masculine and feminine) | يفعلون |
| jamsaɦuuna | CVC-CV-CVV-CV | VC-a-VC-UF-a-UF-uu-VC-a | 'they wipe' (plural masculine and feminine) | يمسحون |
| tafealiina | CVC-CV-CVV-CV | UP-a-UF-VC-a-VC-ii-VC-a | 'you do' (singular feminine) | تفعلين |
| tamsaɦiina | CVC-CV-CVV-CV | UP-a-VC-UF-a-UF-ii-VC-a | 'you wipe' (singular feminine) | تمسحين |
| jamsaɦaani | CVC-CV-CVV-CV | VC-a-VC-UF-a-UF-aa-VC-i | 'they wipe' (dual masculine) | يمسحان |
| tamsaɦaani | CVC-CV-CVV-CV | UP-a-VC-UF-a-UF-aa-VC-i | 'they wipe' (dual feminine) | تمسحان |
| walbuɦuuθi | CVC-CV-CVV-CV | VC-a-VC-VP-u-UF-uu-UF-a | 'and research' | والبحوث |
| maktabatan | CVC-CV-CV-CVC | VC-a-UP-UP-a-VP-a-UP-a-VC | 'library' | مكتبة |
| jansahibu | CVC-CV-CV-CV | VC-a-VC-UF-a-UF-i-VP-u | 'he withdraws' | ينسحب |
| jataʃaffaɦu | CV-CV-CVC-CV-CV | VC-a-UP-a-UF-a-UF-UF-a-UF-u | 'he browses' | يتصفح |
| ?aleilmijati | CVC-CVC-CV-CV-CV | UP-a-VC-VC-i-VC-VC-i-VC-a-UP-i | 'the scientific' | العلمية |
| ?addirasaati | CVC-CV-CVV-CVV-CV | UP-a-VP-VP-i-VC-aa-UF-aa-UP-i | 'the studies' | الدراسات |

Appendix B

Spectrograms of the Consonant-Vowel Pairs

The spectrograms of the consonants /d/, /t/, /s/, /k/, and /θ/, and their counterparts the pharyngealised consonants /d̤/, /t̤/, /s̤/, /k̤/, and /θ̤/ with the three vowels /aa/, /uu/, and /ii/, are given in Figures 3.4 to 3.6 in Chapter 3. In this appendix the spectrograms of the rest of all possible consonant-vowel pairs (long vowels) are given in Figures B.1 to B-7.

Figures B-1 and B-2 display the spectrograms of 16 consonants followed by the vowel /aa/. Figures B-3 and B-4 display the spectrograms of 16 consonants followed by the vowel /uu/. Figures B-5 and B-6 display the spectrograms of 16 consonants followed by the vowel /ii/. Figure B-7 displays the spectrograms of the consonant /l/ and its pharyngealised counterpart /l̤/ followed by the three vowels /aa/, /uu/ and /ii/.

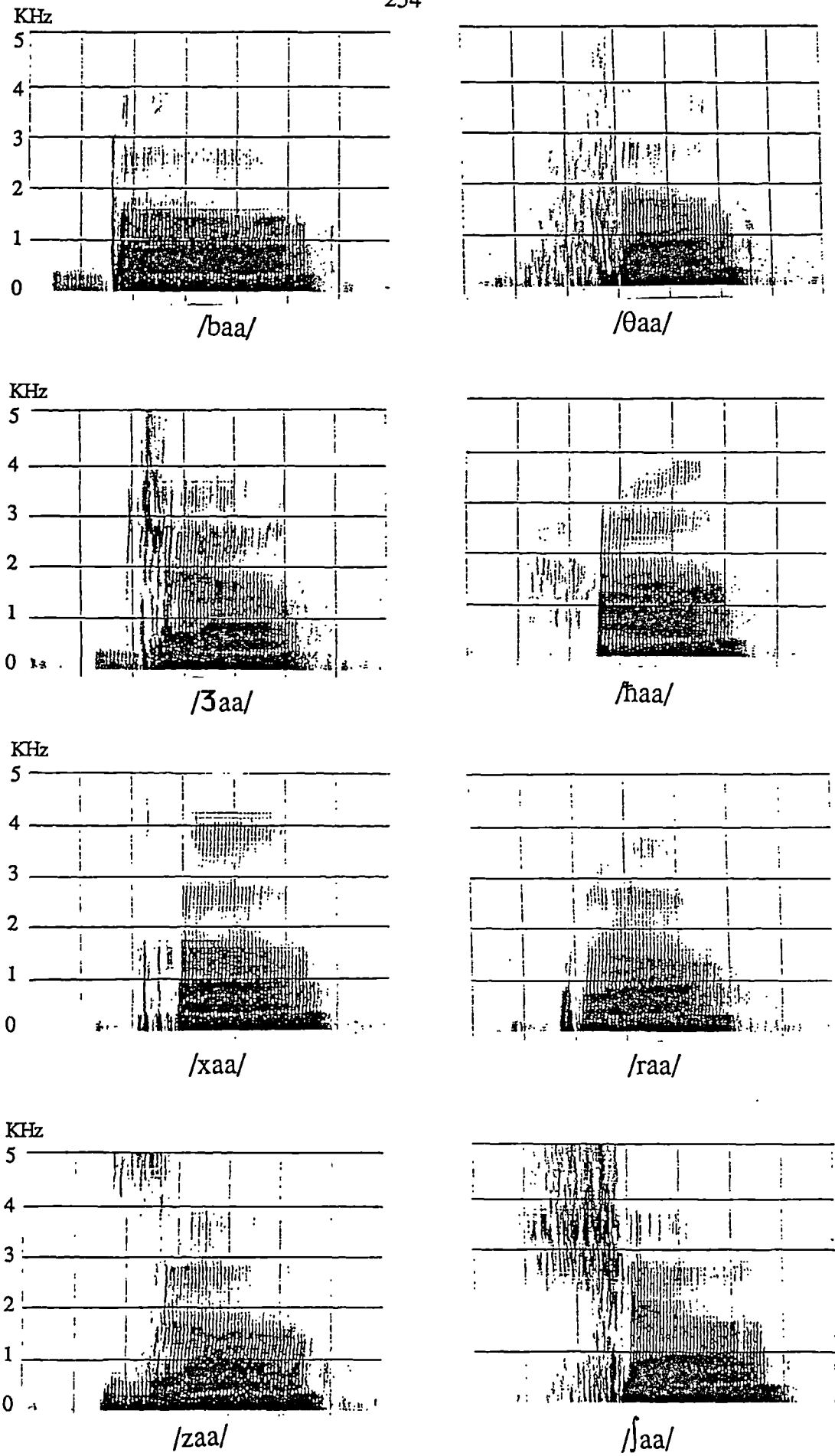


Fig. B.1 Consonant-vowel spectrograms (vowel /aa/)

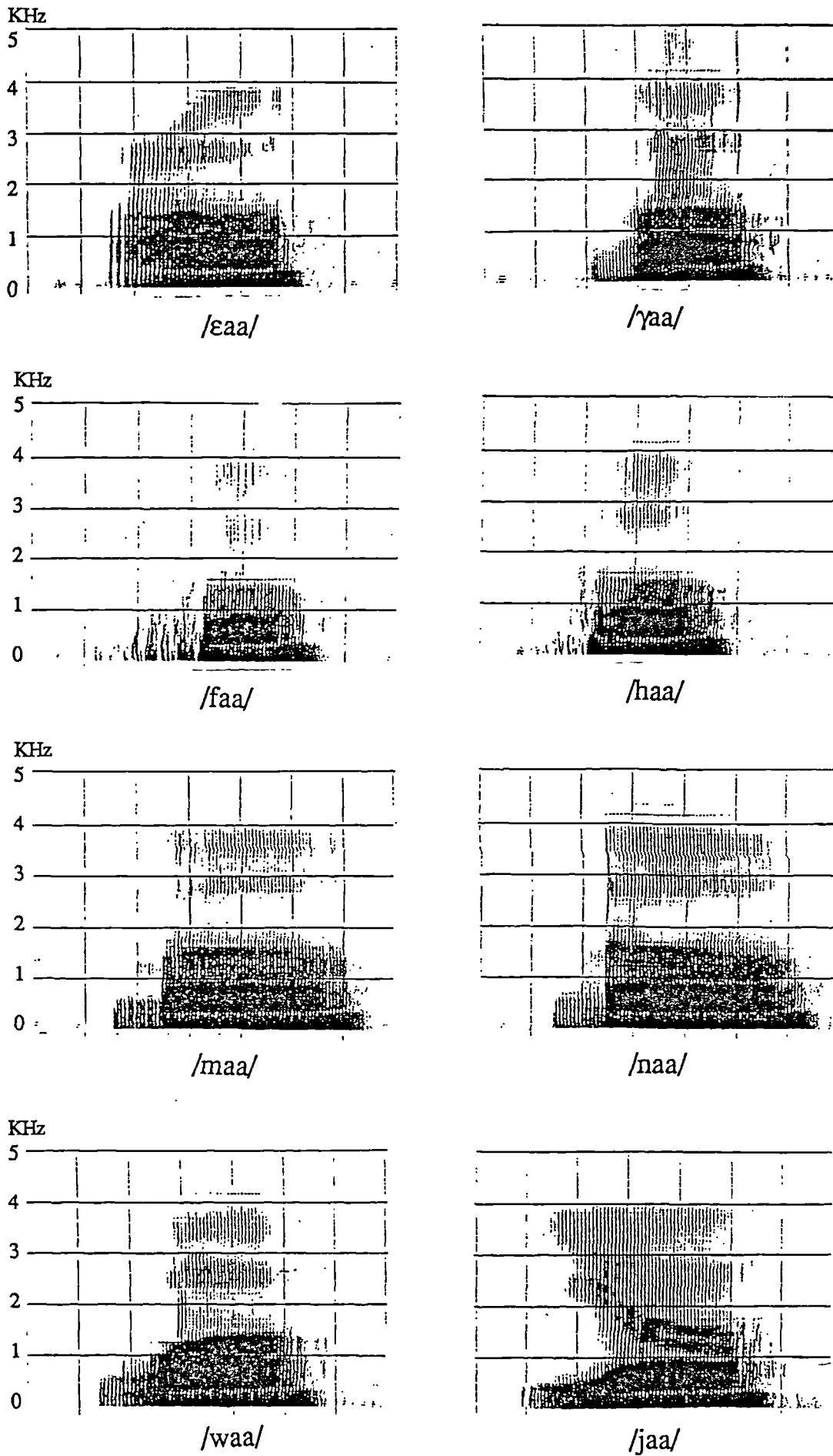


Fig. B.2 Consonant-vowel spectrograms (vowel /aa/)

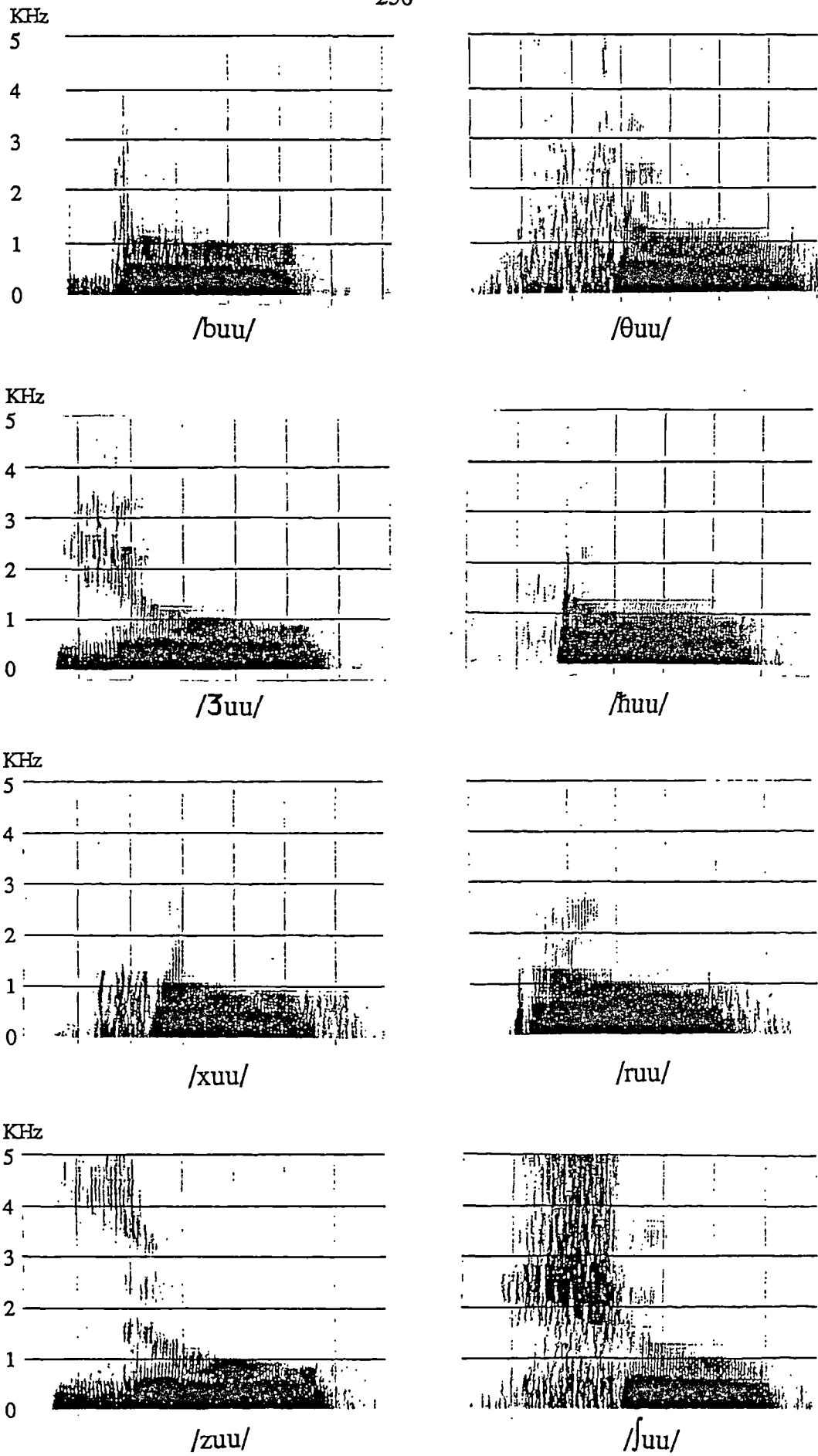


Fig. B.3 Consonant-vowel spectrograms (vowel /uu/)

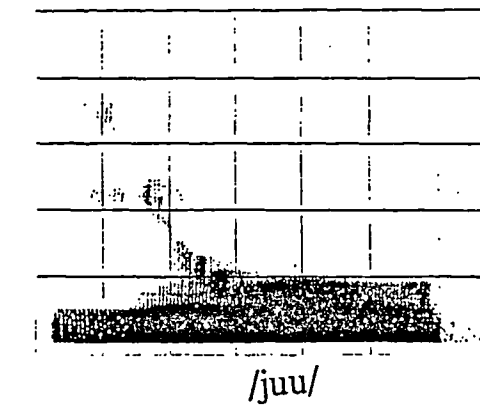
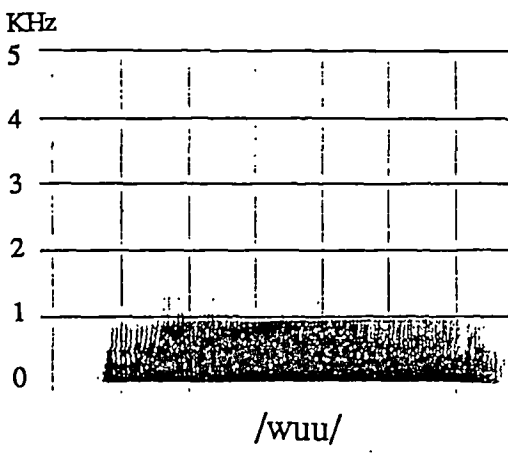
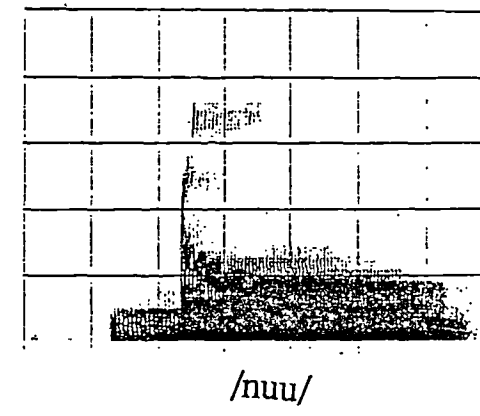
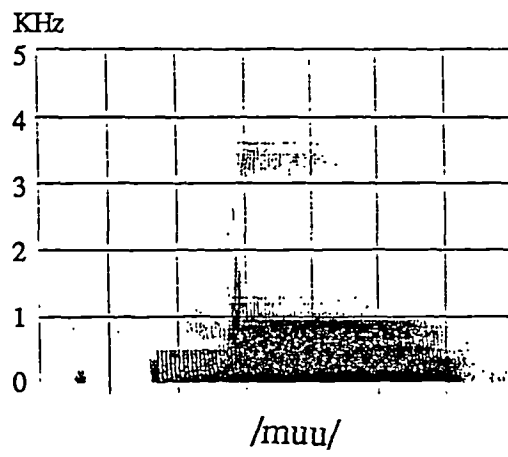
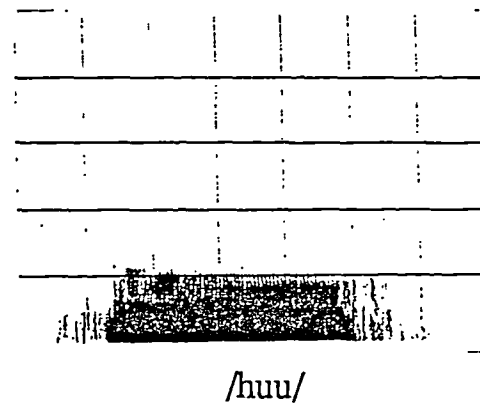
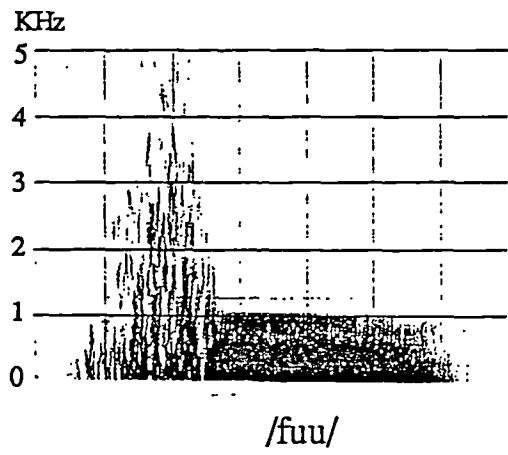
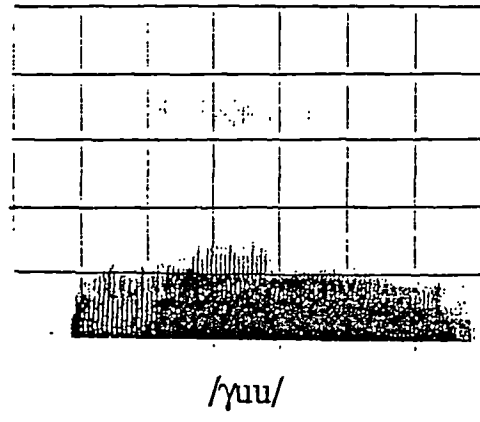
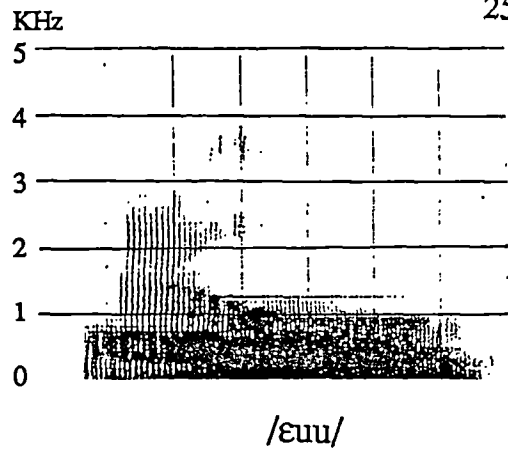


Fig. B.4 Consonant-vowel spectrograms (vowel /uu/)

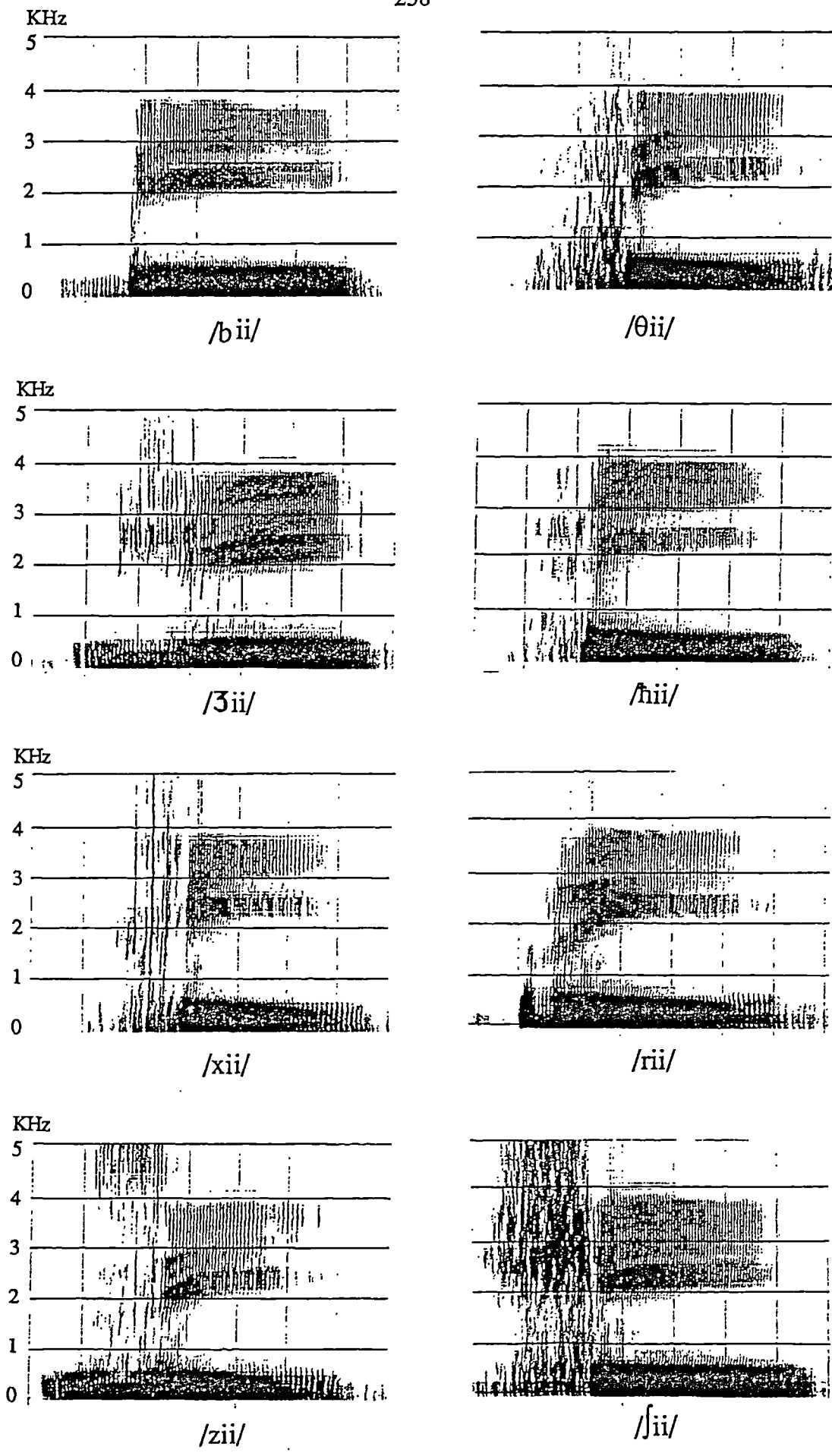


Fig. B.5 Consonant-vowel spectrograms (vowel /ii/)

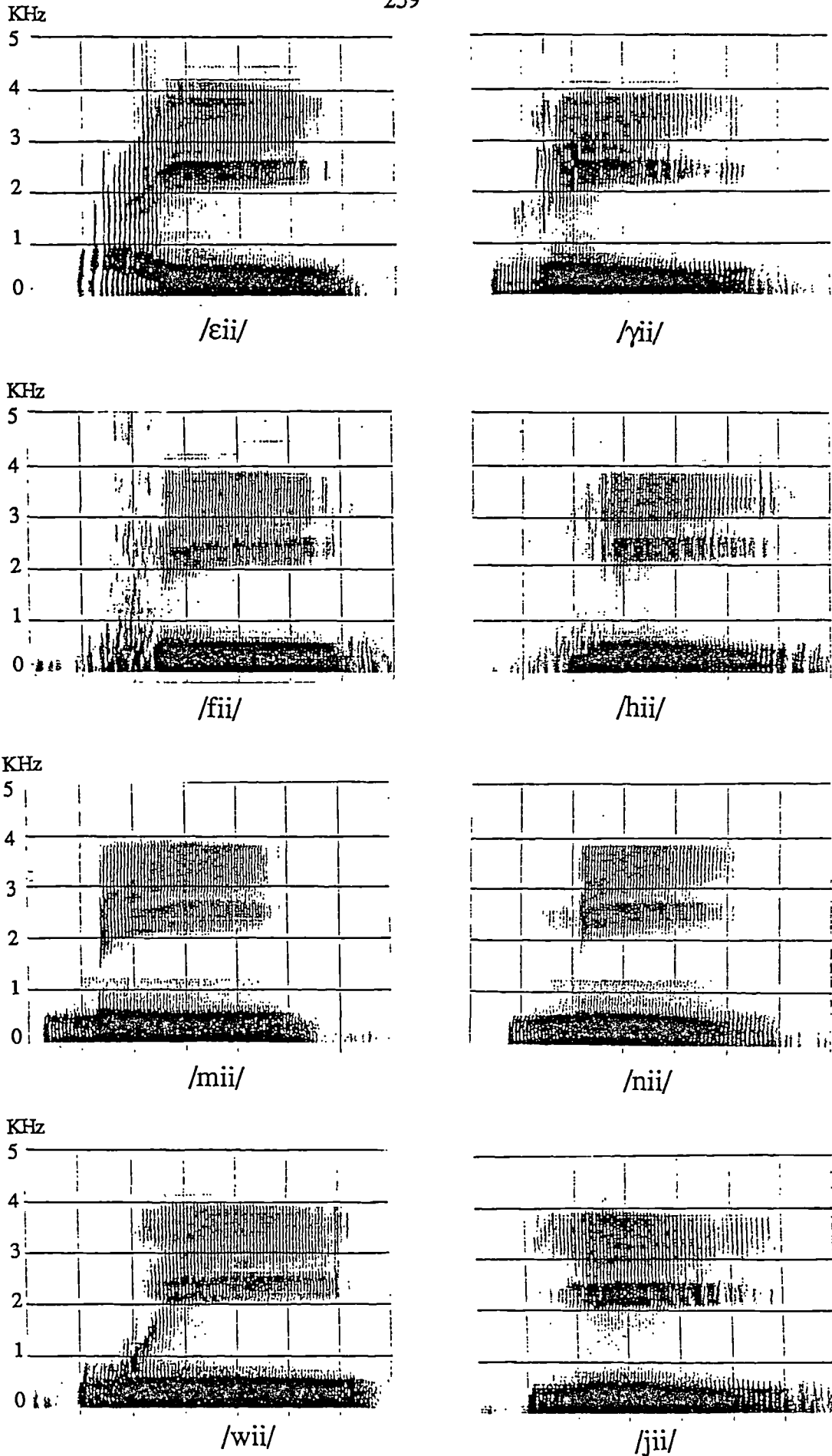


Fig. B.6 Consonant-vowel spectrograms (vowel /ii/)

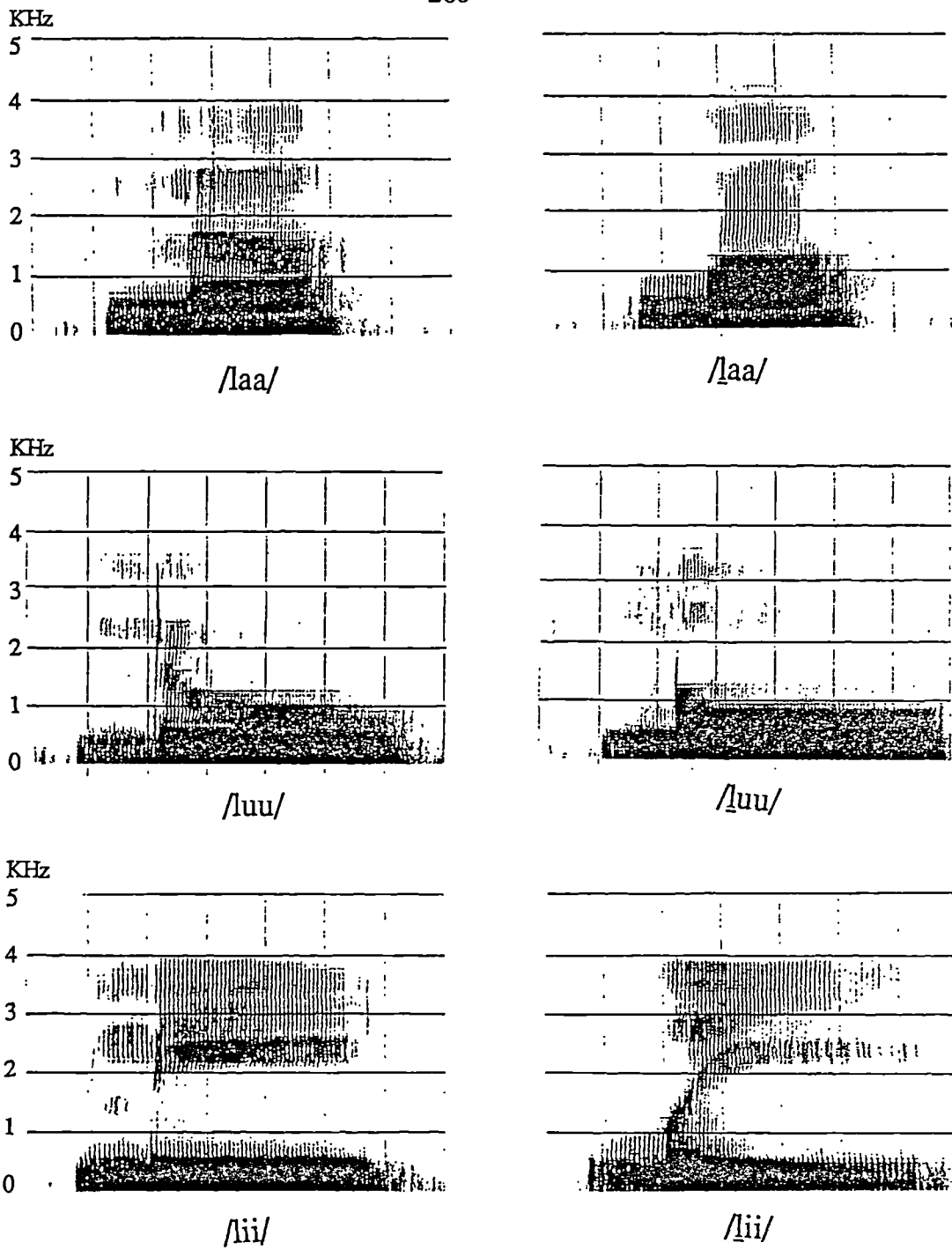


Fig. B.7 Consonant-vowel spectrograms (vowels /aa/, /uu/, and /ii/)

Appendix C

Mel Scale and Critical Bands

C.1 Threshold of Hearing

The absolute sensitivity of the human ear is measured as the smallest sound pressure (SP) which leads to the sensation of hearing. The threshold of hearing depends on the frequency of the sound. Figure C.1 displays this threshold as a function of frequency for a typical young adult [96].

The human ear is most sensitive between 1000 and 3000 Hz with the threshold rising from lower and higher frequencies. If the threshold at 1000 Hz is taken as a reference (see Figure C.1), the signal is to be increased a hundred times to reach the threshold at 100 Hz and 15,000 Hz, and a thousand times to reach the threshold at 18,000 Hz. The threshold of pain occurs more or less uniformly at sound intensity equal to 140 dB.

The frequency limits of hearing are generally considered to lie between 20 and 20,000 Hz.

C.2 Pitch and Mel Scale

Pitch is the subjective attribute of a sound which corresponds to the physical attribute of frequency. Although the pitch of a pure tone is monotonically related to its frequency, a linear relationship does not hold. The unit of pitch is the 'mel'. The mel scale has been constructed on the basis of subjective pitch evaluations. This involves the determination of the frequency corresponding to halving and doubling of the pitch and equal increments of the pitch by naive listeners. A tone with a frequency of 1000 Hz is defined as having a pitch of 1000 mels. A tone with a pitch of 500 mels sounds half as high as one with a pitch of 1000 mels. However, its frequency will be 400 Hz. Similarly a tone with a pitch of 2000 mels will sound twice as high as one with a pitch of 1000 mels, yet its frequency will be 3000 Hz rather than 2000 Hz. Figure C.2 illustrates the relationship between the pitch scale and the frequency scale for pure tone of 40 dB intensity [96].

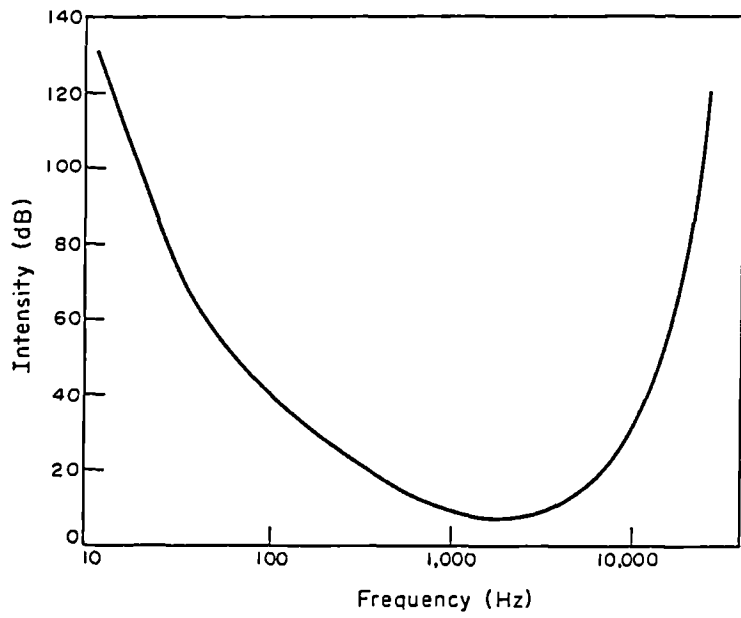


Fig. C.1 Threshold of hearing as a function of frequency

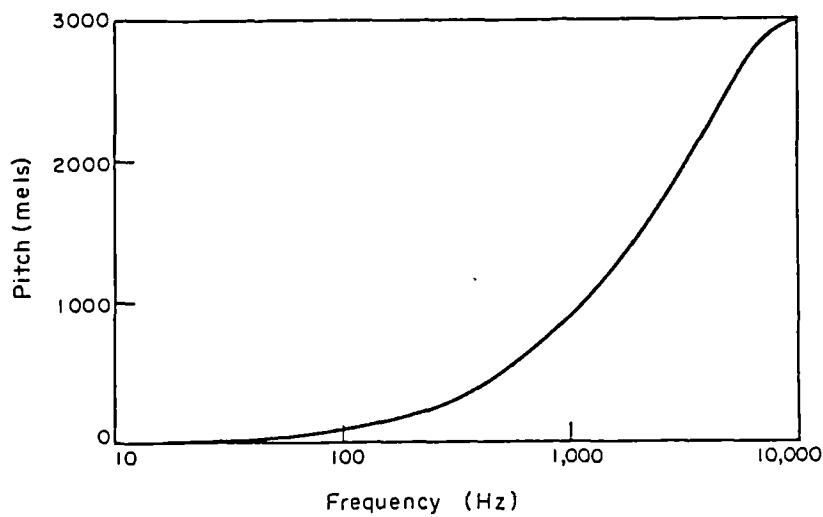


Fig. C.2 Relationship between pitch scale and frequency

The mel scale is essentially linear at low frequencies and logarithmic at higher frequencies. A technically useful approximation to the mel scale is of the form [102]:

$$y = k \log (1 + f / 1000) \quad (\text{C.1})$$

where f is the frequency in Hz, k is a constant. The constant k is computed with the consideration that a tone with a frequency of 1000 Hz is defined as having a pitch of 1000 mels. Thus, k is equal to 3322. A conversion of a frequency to a mel scale is roughly identical with an estimate of the spatial position of the corresponding point of maximum excitation on the basilar membrane in the cochlea (in the inner ear).

The cochlea, a liquid-filled tube located in the inner ear, performs a continuous broad-band analysis of the sound which enters the ear, and transmits the results to the brain through the neural fibre outputs of the cochlea. The basilar membrane in the cochlea, which performs this spectral analysis, has different frequency response along its length. Each location along the basilar membrane has a characteristic frequency, at which it vibrates maximally for a given input sound. For a specific location, the response curve as a function of the vibration frequency at the input of the cochlea is that of a bandpass filter with almost constant Q (fixed ratio of centre frequency to bandwidth). Because of this constant-percentage bandwidth, frequency resolution along the basilar membrane is best at low frequencies. For every input frequency, there is a point on the basilar membrane of maximal vibration.

According to the mel scale, the frequency range over which the human ear is able to perceive sounds can be divided into a bank of bandpass filters. These filters are linearly spaced below 1000 Hz and logarithmically spaced above 1000 Hz. The filters under 1000 Hz have fixed bandwidths and are taken equal to 100 Hz. The filters at and above 1000 Hz follow a logarithmic distribution according to Eq. (C.1), and these filters are assumed to have constant Q which is given as follows:

$$Q = f_0 / \text{BW} \quad (\text{C.2})$$

where f_0 and BW are the filter's centre frequency and bandwidth respectively. From Eq. (C.1), the frequency f is given as a function of its value y at mel scale as follows:

$$f = 10^3 (10^{y/k} - 1) \quad (\text{C.3})$$

For a bandpass filter with $f_0 = 1000$ Hz, $y_0 = 1000$ mels and a bandwidth equals to 100 mels, Q is computed by substituting Eq. (C.3) in Eq. (C.2) to yield $Q = 7.3$. In order to have a flat composite spectrum over the whole frequency range of the filter bank, the centre frequency of a filter i is computed as follows:

$$f_i = f_{i-1} + BW_{i-1} = f_{i-1} (1 + 1/Q) = 1.137 f_{i-1} \quad (\text{C.4})$$

For $f_{i-1} = 1000$ Hz, the centre frequency of the following filter is at 1137 Hz. Table C.1 illustrates values of the centre frequency and bandwidth of a bank of 22 filters covering the range 50-4980 Hz [140], where the centre frequencies above 1000 Hz follow Eq. (C.4).

| Filter No. | Centre Frequency Hz | Bandwidth Hz |
|------------|------------------------|-----------------|
| 1 | 100 | 100 |
| 2 | 200 | 100 |
| 3 | 300 | 100 |
| 4 | 400 | 100 |
| 5 | 500 | 100 |
| 6 | 600 | 100 |
| 7 | 700 | 100 |
| 8 | 800 | 100 |
| 9 | 900 | 100 |
| 10 | 1000 | 118 |
| 11 | 1137 | 146 |
| 12 | 1292 | 166 |
| 13 | 1469 | 189 |
| 14 | 1671 | 215 |
| 15 | 1899 | 244 |
| 16 | 2159 | 278 |
| 17 | 2455 | 316 |
| 18 | 2791 | 359 |
| 19 | 3173 | 408 |
| 20 | 3607 | 464 |
| 21 | 4101 | 527 |
| 22 | 4662 | 599 |

Table C.1 Filter bank centre frequencies and bandwidths (mel scale)

C.3 Critical Bands

When a weak tone is heard in the presence of an adjacent tone, the threshold for hearing the first tone is raised. This phenomenon is known as 'masking'. It was found that the threshold is raised only when the tones are close to each other in the frequency. If they are more than a critical distance apart, the second tone (whose intensity is above the hearing threshold) has no effect on the threshold for hearing the first tone[96]. This has led to the concept of the critical band. Signals within the critical band influence the perception of each other.

Critical bands are measured throughout the frequency range of hearing by listening to tones mixed with band-limited noise. The tone is set at the centre frequency of the band of noise. As the bandwidth of the noise is increased, the intensity at which the tone was just perceived is also increased until the bandwidth of the noise is equal to the critical band. Thereafter, the intensity for hearing the tone remains constant. It has been found that critical bandwidth increases as the centre frequency is raised. The critical bandwidth for a centre frequency of 200 Hz is found to be about 100 Hz, and for 5000 Hz about 1000 Hz [96].

In the cochlea, the point of maximum vibration moves along the basilar membrane as the frequency of excitation is increased. The critical bandwidths correspond approximately to fixed spacings (1.5 mm spacing) along the basilar membrane, suggesting that a set of 24 bandpass filters would model the basilar membrane well. A perceptual measure, called the 'Bark' scale [36] or 'critical-band rate', relates acoustical frequency to perceptual frequency resolution, in which one Bark covers one critical bandwidth over the whole frequency range, and corresponds nearly to a pitch interval of 100 mels. Table C.2 gives the values for preferred frequencies defining the limits of auditory critical bands [36].

An analytical expression [37] mapping the frequency f into critical-band rate Z , and another expression for critical bandwidth CB are given as follows:

$$Z_i = 13 \arctan (0.76 f) + 3.5 \arctan (f / 7.5)^2 \quad (C.5)$$

$$CB_i = 25 + 75 (1 + 1.4 f^2)^{0.69} \quad (C.6)$$

where f is taken in KHz. These expressions approximate the tabulated data with an accuracy of $\pm 10\%$. From Table C.2, we notice that the critical bandwidth is constant at low frequencies but increases with the logarithm of frequency at high frequencies. Also the critical-band rate is proportional to frequency at low frequencies, but at medium and high frequencies it is proportional to the logarithm of frequency. The critical bands have a certain width, but that their position on the frequency scale is not fixed.

| Critical Band Rate Bark | Centre Frequency Hz | Critical Bandwidth Hz |
|-------------------------------|------------------------|-----------------------------|
| 1 | 50 | 100 |
| 2 | 150 | 100 |
| 3 | 250 | 100 |
| 4 | 350 | 100 |
| 5 | 450 | 110 |
| 6 | 570 | 120 |
| 7 | 700 | 140 |
| 8 | 840 | 150 |
| 9 | 1000 | 160 |
| 10 | 1170 | 190 |
| 11 | 1370 | 210 |
| 12 | 1600 | 240 |
| 13 | 1850 | 280 |
| 14 | 2150 | 320 |
| 15 | 2500 | 380 |
| 16 | 2900 | 450 |
| 17 | 3400 | 550 |
| 18 | 4000 | 700 |
| 19 | 4800 | 900 |
| 20 | 5800 | 1100 |
| 21 | 7000 | 1300 |
| 22 | 8500 | 1800 |
| 23 | 10500 | 2500 |
| 24 | 13500 | 3500 |

Table C.2 Values of critical band rate and critical bandwidth as a function of frequency